# Constraining Credences

by

Sarah Moss

A.B., Harvard University (2002)
B.Phil., Oxford University (2004)



Submitted to the Department of Linguistics and Philosophy
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the



MASSACHUSETTS INSTITUTE OF TECHNOLOGY

[June 2009]
April 2009



Author .....
Department of Linguistics and Philosophy
April 27, 2009

Certified by .....
Robert C. Stalnaker
Laurence Rockefeller Professor of Philosophy
Thesis Supervisor

Accepted by .....
Alex Byrne
Professor of Philosophy
Chair of the Committee on Graduate Students

**Constraining Credences**

by

Sarah Moss

Submitted to the Department of Linguistics and Philosophy
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy.

April 27, 2009

This dissertation is about ways in which our rational credences are constrained: by norms governing our opinions about counterfactuals, by the opinions of other agents, and by our own previous opinions.

In Chapter 1, I discuss ordinary language judgments about sequences of counterfactuals, and then discuss intuitions about norms governing our credence in counterfactuals. I argue that in both cases, a good theory of our judgments calls for a static semantics on which counterfactuals have substantive truth conditions, such as the variably strict conditional semantic theories given in STALNAKER 1968 and LEWIS 1973a. In particular, I demonstrate that given plausible assumptions, norms governing our credences about objective chances entail intuitive norms governing our opinions about counterfactuals. I argue that my pragmatic accounts of our intuitions dominate semantic theories given by VON FINTEL 2001, GILLIES 2007, and EDGINGTON 2008.

In Chapter 2, I state constraints on what credence constitutes a perfect compromise between agents who have different credences in a proposition. It is sometimes taken for granted that disagreeing agents achieve a perfect compromise by splitting the difference in their credences. In this chapter, I develop and defend an alternative strategy for perfect compromise, according to which agents perfectly compromise by coordinating on the credences that they collectively most prefer, given their purely epistemic values.

In Chapter 3, I say how your past credences should constrain your present credences. In particular, I develop a procedure for rationally updating your

credences in *de se* propositions, or sets of centered worlds. I argue that in forming an updated credence distribution, you must first use information you recall from your previous self to form a hypothetical credence distribution, and then change this hypothetical distribution to reflect information you have genuinely learned as time has passed. In making this proposal precise, I argue that your recalling information from your previous self resembles a familiar process: agents' gaining information from each other through ordinary communication.

## Acknowledgements

First and foremost, thanks to my committee: Bob Stalnaker, Steve Yablo, and Roger White. They have provided me with generous comments on all the ideas in my dissertation, as well as many others that did not make the cut. I will never forget afternoons on which I stayed in Bob's office until he had to get up to turn on the light as the sun set outside. I am grateful not just for my committee's comments, but also for their encouragement and wise advice, both always given at just the right time.

Thanks to audiences at NYU, Berkeley, Michigan, MBPA, Princeton-Rutgers graduate conference 2007, MITing of the Minds 2008, BSPC 2008, Formal Epistemology Workshop 2008, First Formal Epistemology Festival, and the 2009 Pacific APA. Thanks to Alex Byrne, Dorothy Edgington, Andy Egan, Allan Gibbard, Sally Haslanger, Ned Hall, Irene Heim, Richard Holton, Jim Joyce, Rae Langton, David Manley, Vann McGee, and Agustin Rayo for helpful discussions and moral support. Thanks to Kai von Fintel for careful mentorship in linguistics, to John Hawthorne for fun brainstorming sessions, and to Tim Williamson for directing me towards the kind and character of philosophy to which I am now dedicated.

Thanks to Ofra Magidor for seven years of helpful phone conversations. Thanks to MIT graduate students of 2004–2009, and especially to Rachael Briggs, Alejandro Perez-Carballo, and Paolo Santorio for feedback on Chapter 3. Thanks to Susanna Rinard for her contagious enthusiasm. Thanks to Seth Yalcin for making my day with productive conversations throughout my first few years of grad school. And thanks to Dilip Ninan for being a role model; I can only aspire to have his philosophical thoughtfulness and patience.

Thanks to my sister for her sense of perspective, and to my mother for her constant confidence in me. Thanks to my father: as in all things, my hope is that I would have made him proud.

And finally, thanks to Eric Swanson: for his attention to broad brush strokes and his patience with detail, and for the unwavering strength and love with which he supports my work, our family, and myself.

## Contents

## 1 Credences in counterfactuals

STALNAKER 1968 and LEWIS 1973a advocate a certain semantics for counterfactuals, conditionals such as:

> If Sophie had gone to the New York Mets parade, she would have seen Pedro Martínez.

Until recently, theirs was the standard theory. But VON FINTEL 2001 and GILLIES 2007 present a problem for the standard semantics: they claim that it fails to explain the infelicity of certain sequences of counterfactuals, namely *reverse Sobel sequences*. Both von Fintel and Gillies propose alternative dynamic semantic theories that explain the infelicity of reverse Sobel sequences, and argue that we should trade in the standard semantics of counterfactuals for theirs.

In the first part of this chapter, I argue that we can and should explain the infelicity of reverse Sobel sequences without giving up the standard semantics. In §1, I present the Stalnaker-Lewis semantics. In §2, I introduce reverse Sobel sequences, discuss the von Fintel and Gillies theories, and say how their theories predict the infelicity of reverse Sobel sequences. In §3, I give my own explanation of why reverse Sobel sequences are generally infelicitous. In §4, I argue that compared to the von Fintel and Gillies theories, my theory appeals to principles that are more independently motivated, and gives a better account of our judgments about sequences of counterfactuals. For instance, I argue that some reverse Sobel sequences are *felicitous*, and that my theory gives a successful account of our judgments about these sequences. In §5, I discuss another potential application of my approach: infelicitous sequences containing 'might' counterfactuals.

In the second part of this chapter, I defend truth-conditional theories of counterfactuals from an objection that is similar in shape but broader in scope. Recently, EDGINGTON 2008 has argued that traditional truth-conditional theories

of counterfactuals cannot account for our judgments about conditionals such as:

> If I had flipped this coin, it would have landed heads.

For instance, Edgington argues that any semantics assigning truth conditions to counterfactuals will have a hard time accounting for how counterfactuals with objectively chancy consequents embed under certain attitude verbs. She concludes that we should adopt a non-truth-conditional theory of counterfactuals. I introduce Edgington's argument in §6. In §7, I develop her concern, saying how the semantic theories advocated by Lewis and Stalnaker appear to have a hard time accounting for embedding data. I have two main aims in discussing this apparent problem for Lewis and Stalnaker. First, in §7, I argue that *pace* Edgington, the challenging judgments are in fact compatible with a truth-conditional semantics for counterfactuals. I discuss several ways in which a truth-conditional theory could account for the embedding data. Then, in §8, I argue that embedding data actually provide an argument for developing a truth-conditional theory of counterfactuals. Ordinary speakers endorse complex norms governing their credence in counterfactuals, and developing a truth-conditional theory of counterfactuals allows us to explain why ordinary speakers endorse just the norms that they do.

## 1.1 Sobel sequences and the standard semantics

Consider the following counterfactual conditional:

(1) If Sophie had gone to the New York Mets parade, she would have seen Pedro Martínez.

Before too much thinking, it is tempting to say that (1) is true just in case all possible worlds in which Sophie goes to the parade are worlds in which she sees Pedro. This is the *strict conditional analysis* of counterfactuals. On this analysis, the context in which a counterfactual is uttered contributes a function to its truth conditions: an accessibility function $f$ from worlds to sets of worlds. 'If $p$, would $q$' then expresses a proposition that is true at a world just in case all the

$p$ worlds that are $f$-accessible from that world are $q$ worlds.[1]

But now consider the following sequence of counterfactuals:

(2a) If Sophie had gone to the parade, she would have seen Pedro.

(2b) But if Sophie had gone to the parade and been stuck behind a tall person, she would not have seen Pedro.

Intuition says that the counterfactuals in (2) can be true together. But the strict conditional analysis predicts otherwise. For on this analysis, (2a) says that all possible worlds in which Sophie goes to the parade are worlds in which she sees Pedro. Given that there are possible worlds in which Sophie goes to the parade and is stuck behind a tall person, this is incompatible with what (2b) says, namely that all possible worlds in which Sophie goes to the parade and is stuck behind a tall person are worlds in which she does not see Pedro. So the strict conditional analysis predicts that (2a) and (2b) cannot be true together.

Sequences like (2) are *Sobel sequences*.[2] LEWIS 1973a made Sobel sequences famous, and was motivated by them to reject the strict conditional analysis of counterfactuals. On both the Lewis analysis and its cousin in STALNAKER 1968, the context in which a counterfactual is uttered contributes a similarity ordering $O$ on worlds to its truth conditions, rather than contributing a function on worlds. Roughly speaking, 'if $p$, would $q$' expresses a proposition that is true at a world just in case all the $p$ worlds closest-by-$O$ to that world are $q$ worlds.[3] Stalnaker and Lewis predict that the counterfactuals in (2) can both be true. For according to them, (2a) says that the closest worlds in which Sophie goes to the parade are worlds in which she sees Pedro, and that is perfectly compatible with what (2b) says, namely that the closest worlds in which Sophie goes to the

1. When I am sure it will not cause any confusion, I will cut corners to make claims more readable, e.g. where '$p$' is a schematic letter to be replaced by a sentence, I use '$p$ worlds' to refer to worlds where the semantic value of that sentence as uttered in the understood context is true.
2. LEWIS 1973a thanks J. Howard Sobel for bringing sequences like (2) to his attention.
3. More precisely, LEWIS 1973a says that 'if $p$, would $q$' expresses a proposition that is non-vacuously true at a world just in case some $p$-and-$q$ world is closer to that world than any $p$-and-not-$q$ world. STALNAKER 1968 says that 'if $p$, would $q$' expresses a proposition that is true at a world just in case the closest $p$ world is a $q$ world. I will not focus on the details of these rival versions of the standard semantics, but I will flag differences between the accounts where they are relevant to my arguments.

parade and is stuck behind a tall person are worlds in which she does not see Pedro. Hence this analysis looks promising, and until recently, most theorists accepted some version of this analysis of counterfactuals.

### 1.2 Reverse Sobel sequences

But VON FINTEL 2001 and GILLIES 2007 raise a problem for the standard analysis of counterfactuals.[4] Suppose we reverse the order of the sentences in (2) to make the following sequence (3):

(3a) If Sophie had gone to the parade and been stuck behind a tall person, she would not have seen Pedro.

(3b) #But if she had gone to the parade, she would have seen Pedro.

Both von Fintel and Gillies say that when uttered in this order, if (3a) is true then (3b) is not true. But according to von Fintel and Gillies, the standard analysis predicts otherwise. For on the standard analysis, the order in which the counterfactuals in (2) and (3) are uttered makes no difference to their semantic value. Even if you have just said that the closest worlds in which Sophie goes to the parade and is stuck behind a tall person are ones where she does not see Pedro, you may go on to truly say that the closest worlds in which she goes to the parade are ones where she sees Pedro, with no fear of contradiction. So the standard analysis predicts that even when uttered in sequence (3), the counterfactual (3b) can be true.

Sequences like (3) are *reverse Sobel sequences*. These sequences motivate von Fintel and Gillies to trade in the standard analysis of counterfactuals for another theory. Surprisingly, they trade in the standard analysis for a variant on the original strict conditional analysis of counterfactuals.[5] In other words, they want to preserve the original claim that 'if $p$, would $q$' is true just in case all the $p$ worlds in a contextually determined set are $q$ worlds. But they augment this claim with a strong claim about the dynamics of conversation: as part of their meaning, counterfactuals effect changes in the context. In particular,

4. In (2001), von Fintel credits Irene Heim with the origination of reverse Sobel sequences.

5. Other proponents of a return to the strict conditional analysis include WARMBROD 1981 and LOWE 1995.

counterfactuals impose demands on the contextually determined domain that subsequent counterfactuals quantify over.[6]

In (2001), von Fintel endorses much of the strict conditional analysis. He adopts the claim that context contributes an accessibility function $f$ to the truth conditions of counterfactuals, a function from worlds to sets of worlds. He adopts the claim that 'if $p$, would $q$' expresses a proposition that is true at a world just in case all the $p$ worlds that are $f$-accessible from that world are $q$ worlds. But von Fintel adds that there is a second contextual parameter relevant to the interpretation of counterfactuals: a similarity ordering on worlds. He also adds that there is another component to the meaning of a counterfactual: its effect on the accessibility function $f$. In particular, 'if $p$, would $q$' demands that from every world, there be some $f$-accessible $p$ worlds.

More precisely, von Fintel says that counterfactuals "update $f$ by adding to it for any world $w$ the closest antecedent worlds" (20). Suppose that the accessibility function of some context maps some world $w$ to a set that contains no $p$ worlds. Uttering 'if $p$, would $q$' in that context updates the accessibility function, so that it maps that same world $w$ to a set that does contain $p$ worlds. In particular, the updated function maps $w$ to the set of all the worlds at least as close to $w$ as the nearest $p$ worlds, by the contextually determined similarity ordering.

I have said that according to von Fintel, 'if $p$, would $q$' *demands* that from every world, there be some $f$-accessible $p$ worlds. There are several ways to understand the nature of this demand. On one version of the proposal, the semantic value of 'if $p$, would $q$' is a pair of update rules: a rule for updating the context set and a rule for updating the accessibility function $f$. On another version, it is part of the meaning of 'if $p$, would $q$' that it presupposes that from every world, there are $f$-accessible $p$ worlds. On either version, the upshot is the same: once 'if $p$, would $q$' is asserted, there must be $p$ worlds in the domain that the counterfactual quantifies over.[7]

6. Strictly speaking, the accessibility function maps each world to its own domain of accessible worlds. The truth of a counterfactual at a world depends on properties of the worlds accessible from that world. So in a sense, a counterfactual quantifies over many domains: one for each world. I trust the reader to read my claims accordingly.

7. What makes a semantic theory dynamic is controversial. Some may prefer to reserve the term 'dynamic' for the first version of von Fintel's proposal. I follow GILLIES 2007 in applying

This dynamic analysis predicts that (3b) must be false, while (2b) can be true. (3a) demands that there be some accessible worlds in which Sophie goes to the parade and is stuck behind a tall person. So once we meet the demands of (3a), there are some accessible worlds in which Sophie goes to the parade and does not see Pedro. (3b) says that Sophie sees Pedro in all accessible worlds in which she goes to the parade. So once we utter (3a) and accommodate its demands, (3b) must be false. But Sobel sequences do not crash in the way that reverse Sobel sequences do. (2a) demands merely that there are some accessible worlds in which Sophie goes to the parade. (2b) says that in all accessible worlds in which Sophie goes to the parade and is stuck behind a tall person, she does not see Pedro. So even once we utter (2a) and accommodate its demands, (2b) can be true.

One might prefer a slight variation on this analysis. In his (1997), von Fintel argues that 'if $p$, would $q$' presupposes a local application of the law of conditional excluded middle: that either all or none of the accessible $p$ worlds are $q$ worlds. If you accept this claim, and also accept that a counterfactual lacks a truth value when this particular presupposition is false, then given the dynamic semantics presented above, you will conclude that (3b) is not false, but merely lacks a truth value. My arguments against the dynamic approach apply equally to this analysis of reverse Sobel sequences.

Gillies 2007 develops a dynamic semantics of counterfactuals similar to von Fintel's. On von Fintel's analysis, there are two contextual parameters: a regularly updated accessibility function, and a similarity ordering on worlds. Gillies posits only one parameter: a *counterfactual hyperdomain*, i.e. a collection of nested sets of worlds. He says that 'if $p$, would $q$' is true just in case all the $p$ worlds in the smallest set in the counterfactual hyperdomain are $q$ worlds. Gillies then adds another component to the meaning of a counterfactual: 'if $p$, would $q$' demands that there are some $p$ worlds in the smallest set in the counterfactual hyperdomain.

Gillies predicts that (3b) cannot be true if (3a) is true, in almost exactly the same way von Fintel does. Once we accommodate the demands of (3a), there are some worlds in the smallest set in the counterfactual hyperdomain in which Sophie goes to the parade and does not see Pedro. (3b) says that Sophie sees

'dynamic' to theories resembling either version.

Pedro in all the worlds in the smallest set in the counterfactual hyperdomain. So once we utter (3a) and accommodate its demands, (3b) cannot be true. GILLIES 2007 concludes that reverse Sobel sequences are *inconsistent*: a reverse Sobel sequence "cannot be interpreted without collapse into absurdity" (28). But the demands of (2a) are weaker, so even once we accommodate them, (2b) can be true.

Besides VON FINTEL 2001 and GILLIES 2007, I know of only one other analysis of conditionals that aims to account for phenomena like the infelicity of (3). WILLIAMS 2006a observes that the indicative analog of (2) is felicitous:

(2a′) If Sophie went to the parade, she saw Pedro.

(2b′) But if Sophie went to the parade and got stuck behind a tall person, she did not see Pedro.

Meanwhile, the indicative analog of (3) is infelicitous:

(3a′) If Sophie went to the parade and got stuck behind a tall person, she did not see Pedro.

(3b′) #But if Sophie went to the parade, she saw Pedro.

Williams accounts for these data by adopting a variant of the strict conditional analysis for indicative conditionals, according to which the domain of the necessity modal is the *context set*: the set of worlds compatible with what is treated as true for purposes of conversation. He says that 'if $p$, $q$' is true just in case all $p$ worlds in the context set are $q$ worlds. Like von Fintel and Gillies, Williams then adds another component to the meaning of a conditional: Williams says that 'if $p$, $q$' presupposes that the context set contains some $p$ worlds.

It is not clear how to generalize Williams' theory to an analysis of counterfactuals. It is okay to utter (1) even if you know that Sophie did not go to the parade. In general, it is okay to utter a counterfactual even if the antecedent is presupposed to be false. So the *counterfactual* conditional 'if $p$, would $q$' does not presuppose that the context set contains some $p$ worlds. Williams does not spell out an analysis of counterfactuals. But he says that the analysis of counterfactuals in GILLIES 2007, though developed independently, is "similar in spirit" to his own theory. Insofar as the generalization of Williams' theory to counter-

factuals resembles the analysis in GILLIES 2007, the concerns I raise for Gillies apply to Williams too.

To sum up how things stand so far: the strict conditional analysis of counterfactuals says that 'if $p$, would $q$' is true just in case all the possible $p$ worlds are $q$ worlds. Sobel sequences motivate Lewis to reject this story for an analysis according to which 'if $p$, would $q$' is true just in case all the closest possible $p$ worlds are $q$ worlds. Reverse Sobel sequences motivate von Fintel and Gillies to reject this story for a variant of the strict conditional analysis, according to which 'if $p$, would $q$' is true just in case all the $p$ worlds in a certain contextually determined domain are $q$ worlds, and the same sentence demands that there be some $p$ worlds in that domain.

Presented with these theories, one might be tempted to revive the standard analysis. Strictly speaking, the standard analysis can accommodate the infelicity of reverse Sobel sequences. The second sentence of a reverse Sobel sequence might be false because the similarity ordering determined by context changes when you utter the first sentence. The advocate of the standard analysis can say that uttering (3a) changes the contextually determined similarity ordering, expanding the set of closest worlds in which Sophie goes to the parade until it includes some worlds in which Sophie is stuck behind a tall person. (3b) would be false as uttered after such a context change.

Of course, it is not exactly in the spirit of the standard analysis to think that it is so easy to change the contextually determined similarity ordering. Lewis and Stalnaker explain why one can truly utter (2b) after (2a) without saying that the contextually determined similarity ordering changes in (2). They simply say that according to the single similarity ordering in play throughout (2), worlds where Sophie is stuck behind a tall person are farther away than some other worlds where she goes to the parade. Given that Lewis and Stalnaker do not posit changes in the similarity ordering to explain (2), it seems against the spirit of the standard analysis to posit such changes to explain (3).

But at this point in the game, von Fintel and Gillies can claim a greater advantage over the standard semantics: the dynamic approach is a stronger theory, yielding systematic predictions about when counterfactuals are felicitous. For example, it is part of the dynamic semantic value of (3a) that it effects particular changes on the domain that counterfactuals quantify over. So the

dynamic theory itself entails that (3b) will be infelicitous after (3a) is uttered. Lewis and Stalnaker may say that (3b) is infelicitous in contexts such that Sophie gets stuck behind a tall person in some of the closest possible worlds in which she goes to the parade. But nothing in their theory predicts that uttering (3a) will make the context be this way. LEWIS 1973a in fact dismisses a version of the *strict conditional analysis* on similar grounds. Lewis considers only judgments about Sobel sequences, not reverse Sobel sequences. He says that appealing to context shifting in order to explain the felicity of Sobel sequences is "defeatist...consign[ing] to the wastebasket of contextually resolved vagueness something much more amenable to systematic analysis than most of the rest of the mess in that wastebasket" (13).

GILLIES 2007 responds to Lewis:

> To see that this kind of story is not the stuff of defeatism we only have to see that the interaction between context and semantic value, mediated by a mechanism of local accommodation, can be the stuff of formal and systematic analysis. To see that this is not a mere loophole, we only have to see that facts about counterfactuals in context—the discourse dynamics surrounding them—are best got at by the kind of story I want to tell. (2)

To sum up: Gillies and von Fintel claim that positing changes in the similarity ordering is the only way for the standard semantics to account for the infelicity of reverse Sobel sequences. If that were true, then on behalf of advocates of the standard semantics, I would concede: we should prefer a theory that yields systematic predictions about when counterfactuals are felicitous. The point of the dynamic approach is to provide just this kind of theory.

## 1.3 A pragmatic account of reverse Sobel sequences

However, I think von Fintel and Gillies are wrong: the standard semantics can account for the infelicity of reverse Sobel sequences, without positing changes in the similarity ordering. In this section, I will give an alternative explanation for the infelicity of reverse Sobel sequences. My explanation is compatible with a Stalnaker-Lewis analysis on which uttering sequences like (2) and (3) does not change the contextually determined similarity ordering.

Suppose we are enjoying a perfectly normal day at the zoo, looking at an

animal in the zebra cage that seems to have natural black and white stripes. It has not recently crossed our minds that the zoo may be running a really low-budget operation, where they paint mules to look like zebras. In this situation, I might have reason to say:

(4a) That animal was born with stripes.

If you are in a slightly pedantic mood, you might reply with the following:

(4b) But cleverly disguised mules are not born with stripes.

This reply may be a non sequitur, perhaps even a little annoying. But otherwise, there is nothing wrong with your reply. On the other hand, once you have mentioned cleverly disguised mules, I would not be willing to repeat my original assertion. I may even feel as if I ought to take back what I said. In other words, there is a contrast between sequence (4) and the following sequence:

(5a) Cleverly disguised mules are not born with stripes.

(5b) #But that animal was born with stripes.

I would like to suggest that Sobel sequences are okay for the same reason (4) is, and reverse Sobel sequences are bad for the same reason (5) is.[8]

So why is (5) bad, while (4) is okay? Here is one intuitive answer: in the above scenario, (5b) is infelicitous because (5a) raises the possibility that the caged animal is a cleverly disguised mule, and the speaker of (5b) cannot rule out this possibility. So (5b) is infelicitous because in the above scenario, it is an epistemically irresponsible thing to say. Meanwhile, it is perfectly okay to utter

8. I discuss a sequence about zebras because zebra examples are familiar, and so using a zebra example is a quick and reliable way to situate my theory among familiar debates. However, our familiarity with zebra examples can create unwanted noise in our judgments about them. Some informants judge there to be a more marked difference between the following conversations, as uttered in New York City:

(4a′) My car is around the corner.

(4b′) But cars get stolen in New York City all the time.

(5a′) Cars get stolen in New York City all the time.

(5b′) #But my car is around the corner.

the same sentences in the reverse order, since uttering (4a) is not epistemically irresponsible when no recherché possibilities are salient, and (4a) does not raise possibilities to salience that the speaker of (4b) irresponsibly ignores.

Our intuitions about (5) point towards a general principle governing assertability.[9]

(EI) It is epistemically irresponsible to utter sentence $S$ in context $C$ if there is some proposition $\phi$ and possibility $\mu$ such that when the speaker utters $S$:

(i) $S$ expresses $\phi$ in $C$
(ii) $\phi$ is incompatible with $\mu$
(iii) $\mu$ is a salient possibility
(iv) the speaker of $S$ cannot rule out $\mu$.

(EI) tells us that if a speaker cannot rule out a possibility made salient by some utterance, then it is irresponsible of her to assert a proposition incompatible with this possibility.[10] Hence we can use (EI) to explain why it is infelicitous to utter (5b) in the scenario described above. It simply remains to be shown that we can use this independently motivated principle to explain why it is generally infelicitous to utter reverse Sobel sequences.

Earlier I stipulated that the speaker of (5b) could not rule out that a certain animal was a cleverly disguised mule. One can make a similar claim about reverse Sobel sequence scenarios: the speaker of the second sentence of a reverse Sobel sequence generally cannot rule out certain possibilities incompatible with the content of her utterance. Given (EI), this explains why it is generally infelicitous to utter the second sentence of a reverse Sobel sequence.

For example, consider again the reverse Sobel sequence:

(3a) If Sophie had gone to the parade and been stuck behind a tall person, she would not have seen Pedro.

(3b) #But if Sophie had gone to the parade, she would have seen Pedro.

9. One might aim to derive this principle from others, e.g. from the knowledge norm of assertion and the principle that a speaker knows a proposition only if she can rule out salient possibilities incompatible with that proposition.
10. Here I am taking possibilities to be propositions.

Here is one way (EI) could explain why it is generally infelicitous to utter (3b) after (3a). Someone who utters (3b) generally will not be able to rule out the possibility that if Sophie had gone to the parade, she might have been stuck behind a tall person. Once someone utters (3a), we may generally reason: "if the speaker of (3a) could rule out the possibility that Sophie might have been stuck behind a tall person if she'd gone to the parade, why would she even bother to talk about what would have happened if she had gone and been stuck? She would have no practical reason to discuss the matter. So since she is discussing the matter, she must not be able to rule out the possibility that Sophie might have been stuck behind a tall person if she'd gone." If the same speaker utters (3b), we may immediately infer that the speaker of (3b) cannot rule out the possibility that Sophie might have been stuck if she had gone to the parade. If a second speaker utters (3b), then we may generally infer that the second speaker also does not rule out the possibility that Sophie might have been stuck if she had gone, since otherwise the second speaker would have corrected the first speaker after hearing her take this possibility seriously.

Given this inference, (EI) entails that it is epistemically irresponsible to utter (3b), since:

(i) (3b) expresses the proposition that Sophie would have seen Pedro if she had gone to the parade.

(ii) The proposition that Sophie would have seen Pedro if she had gone to the parade is incompatible with the possibility that Sophie might have been stuck behind a tall person if she had gone to the parade.

(iii) The possibility that Sophie might have been stuck behind a tall person if she had gone to the parade is a salient possibility.

(iv) The speaker of (3b), at the time at which she utters (3b), cannot rule out the possibility that Sophie might have been stuck behind a tall person if she had gone to the parade.

The same goes for other reverse Sobel sequences. Here is one way to apply (EI) to a reverse Sobel sequence case: 'if $p$ and $r$, would not-$q$' raises a certain possibility to salience, namely that $r$ might have been the case if $p$ had been the case. It is not hard to raise this possibility to salience; sometimes merely mentioning the possibility that $r$ suffices. Furthermore, the speaker who then utters

'if $p$, would $q$' generally cannot rule out this possibility. Finally, the speaker of 'if $p$, would $q$' expresses a proposition incompatible with this possibility. For this reason, it is generally infelicitous to utter the second sentence of a reverse Sobel sequence: it is epistemically irresponsible to assert a proposition incompatible with an uneliminated possibility that the first sentence raises to salience.[11]

This way of applying (EI) to a reverse Sobel sequence case depends on the following: that 'if $p$, would $q$' expresses a proposition incompatible with the possibility that $r$ might have been the case if $p$ had been the case. To make this more precise: even once the second sentence of a reverse Sobel sequence is uttered, it is still an accepted background fact that if $p$ and $r$, would not-$q$. In Lewis's logic of counterfactuals, we can derive a contradiction from this proposition, together with the proposition expressed by 'if $p$, would $q$' and the possibility that if $p$, might $r$.

Here are the relevant rules and axioms of **VC**, Lewis's official logic of counterfactuals:[12]

| | |
|---|---|
| rule 2: | *Deduction within conditionals*: for any $n \geq 1$, $\dfrac{\vdash (\chi_1 \wedge \ldots \wedge \chi_n) \supset \psi}{\vdash ((\phi \boxrightarrow \chi_1) \wedge \ldots \wedge (\phi \boxrightarrow \chi_n)) \supset (\phi \boxrightarrow \psi)}$ |
| axiom 1: | *Truth-functional tautologies* |
| axiom 2: | *Definitions of non-primitive operators* |
| axiom 5: | $(\phi \boxrightarrow \neg\psi) \vee (((\phi \wedge \psi) \boxrightarrow \chi) \equiv (\phi \boxrightarrow (\psi \supset \chi)))$ |

Using these rules and axioms, we can derive a contradiction from the proposition expressed by 'if $p$, would $q$' and the salient possibility that if $p$ might $r$, as follows:

| | | |
|---|---|---|
| 1. | $p \diamondrightarrow r$ | salient possibility |
| 2. | $\neg(p \boxrightarrow \neg r)$ | 1, axiom 2 |
| 3. | $(p \wedge r) \boxrightarrow \neg q$ | background facts |
| 4. | $(p \boxrightarrow \neg r) \vee (((p \wedge r) \boxrightarrow \neg q) \equiv (p \boxrightarrow (r \supset \neg q)))$ | axiom 5 |
| 5. | $p \boxrightarrow (r \supset \neg q)$ | 2, 3, 4, axiom 1 |
| 6. | $((r \supset \neg q) \wedge q) \supset \neg r$ | axiom 1 |
| 7. | $((p \boxrightarrow (r \supset \neg q)) \wedge (p \boxrightarrow q)) \supset (p \boxrightarrow \neg r)$ | 6, rule 2 |

11. For simplicity, I talk as if infelicity is a property of utterances. Strictly speaking, infelicity is audience-relative: an utterance sounds infelicitous to an agent insofar as she takes the resultant assertion to be epistemically irresponsible.
12. See Lewis 1973a p.132 for a complete axiomatization for **VC**.

| | | |
|---|---|---|
| 8. | $\neg(p \boxrightarrow q)$ | 2, 5, 7, axiom 1 |
| 9. | $p \boxrightarrow q$ | expressed proposition |
| 10. | $\bot$ | 8, 9, axiom 1 |

The second sentence of a reverse Sobel sequence expresses the proposition that if $p$, would $q$. For example, (3b) expresses the proposition that if Sophie had gone to the parade, she would have seen Pedro. It is important to realize that while (2a) expresses the same proposition, this proposition is generally no longer common ground once (2b) is uttered.

Since the proposition expressed by (2a) is no longer common ground when (2b) is uttered, one cannot derive a contradiction from the proposition expressed by (2b) and the possibility raised by (2b) itself, given what is common ground when (2b) is uttered. By contrast, the proposition expressed by (3a) does remain common ground when (3b) is uttered. So one can derive a contradiction from the proposition expressed by (3b) and the possibility raised by (3a), given what is common ground when (3b) is uttered, as outlined above.

In the zebra examples, there is a similar asymmetry in whether the proposition expressed by the first sentence typically remains common ground throughout the sequence. Once the speaker of (4b) says that cleverly disguised mules are not born with stripes, it is typically no longer common ground that the caged animal under discussion was born with stripes. But even once the speaker of (5b) says that the caged animal was born with stripes, it is typically still common ground that cleverly disguised mules are not born with stripes.

Our natural responses to these examples are independent evidence of this asymmetry. For instance, it is natural to respond to (5b) by saying:

(5c) But how do you know that this animal was born with stripes? After all, we said that mules are not born with stripes, and for all you know, this animal might be a mule.

But an analogous response to (4b) typically sounds bad:

(4c) #But how do you know that mules are *not* born with stripes? After all, we said that this animal was born with stripes, and for all you know, this animal might be a mule.

And the analogous response to (2b) sounds similarly unnatural:

(2c) #But how do you know that if Sophie had gone and been stuck behind a tall person, she would have missed Pedro? After all, we said that if she had gone, she would have seen Pedro, and for all you know, if she had gone, she might have been stuck behind a tall person.

I will not defend any general theory of how the common ground of a conversation behaves under various conversational pressures. Ultimately, what matters for my purposes is not the exact nature of the mechanism at work in these examples. I am simply interested in general arguments concerning whether this mechanism is semantic or pragmatic in nature.

So far I have spelled out one way to derive a contradiction from a salient possibility and the proposition expressed by the second sentence of a reverse Sobel sequence. There are other ways to apply (EI) to a reverse Sobel sequence case. The second step of the above derivation appeals to axiom 2 of **VC**, and in particular to Lewis's definition of the 'might' operator. For Lewis, 'might' and 'would' counterfactuals are duals:

$$(\phi \diamond\!\!\rightarrow \psi) \equiv \neg(\phi \mathbin{\Box\!\!\rightarrow} \neg\psi)$$

However, this duality thesis is a contentious assumption. For now, I wish to remain neutral about the duality of 'might' and 'would' counterfactuals. If you reject the duality thesis, there are other ways to derive a contradiction from a salient possibility and the proposition expressed by 'if $p$, would $q$'. For instance, you might think that the first sentence of a reverse Sobel sequence raises the possibility that it is not the case that if $p$ were the case, then not-$r$ would be the case, even though you reject that this possibility is equivalent to the possibility that if $p$ were the case, $r$ might be the case.

Alternatively, you may be one of many theorists who are motivated to reject the duality thesis in order to accept the law of conditional excluded middle.[13] In other words, you may think that one of the following must hold: that if $p$ were

13. Lewis 1973a demonstrates that the duality thesis and the law of conditional excluded middle together entail the equivalence of 'might' and 'would' counterfactuals. Many have responded to this argument by rejecting the duality thesis; see Williams 2006b, DeRose 1997, 1994, and Heller 1995 for some examples. See Stalnaker 1978 for arguments in favor of the law of conditional excluded middle.

the case, then $r$ would be the case, or that if $p$ were the case, then not-$r$ would be the case. In that case, you will likely think that it does not take much to raise the possibility that the first of these is true, and you will likely accept that 'if $p$ and $r$ were the case, then not-$q$ would be the case' raises the possibility that if $p$ were the case, then $r$ would be the case. Given that there are possible $p$ worlds, it is again possible to derive a contradiction between this salient possibility and the proposition expressed by the second sentence in the reverse Sobel sequence. Simply replace steps 1–2 of the above derivation with the following:

| | | |
|---|---|---|
| 1a. | $p \boxrightarrow r$ | salient possibility |
| 1b. | $(\neg r \wedge r) \supset \bot$ | axiom 1 |
| 1c. | $((p \boxrightarrow \neg r) \wedge (p \boxrightarrow r)) \supset (p \boxrightarrow \bot)$ | 2, rule 2 |
| 1d. | $\neg(p \boxrightarrow \bot)$ | assumption |
| 2. | $\neg(p \boxrightarrow \neg r)$ | 1, 3, 4, axiom 1 |

The proposition that if $p$, would $r$ is stronger than the proposition that if $p$, might $r$. So this alternative derivation proceeds from stronger assumptions. But the derivation does not appeal to the duality of 'might' and 'would' counterfactuals.

To sum up: independently of various semantic assumptions, one can show that reverse Sobel sequence cases fit the conditions stated in (EI). 'If $p$ and $r$, would not-$q$' raises a possibility to salience, and that possibility contradicts the proposition expressed by 'if $p$, would $q$'. Given (EI), this entails that the speaker of the second sentence is epistemically irresponsible. My proposal is that the second sentence of a reverse Sobel sequence is infelicitous because it is an epistemically irresponsible thing to say.

So far I have taken possibilities to be propositions. Instead, you might say that a possibility is a world, and that a possibility is salient in the sense relevant to (EI) simply when it is contained in the context set of a conversation. (EI) would then entail that a speaker is epistemically irresponsible if she asserts a proposition that is false at some possibility in the context set of her conversation, if she cannot rule out that this possibility is actual. You might also say that possibilities are added to the context set as speakers accommodate presuppositions. For instance, you might say that in the zoo context described above, 'cleverly disguised mules are not born with stripes' presupposes that the caged animal in view might be a cleverly disguised mule. You might say that (3a)

presupposes that it might be the case that if Sophie had gone to the parade, she might have been stuck behind a tall person. On this theory, a speaker should not utter (3b) or (5b) because she would thereby express a proposition false at live possibilities that are contained in the context set once she accommodates the presuppositions of (3a) or (5a).[14]

This presupposition-based theory shares a lot with the dynamic accounts discussed in §2. On all these theories, the first sentence of a reverse Sobel sequence affects the context by introducing a demand—roughly, a demand about some possibility—and this causes the second sentence to be infelicitous. But even this presupposition-based variation on my proposal differs from the dynamic accounts. There are technical differences: von Fintel and Gillies say that the trouble-making possibility is a world in which Sophie goes to the parade and gets stuck behind a tall person, whereas on the account just sketched, it is a world in which Sophie gets stuck behind a tall person *in some of the closest worlds in which she goes to the parade.*

Furthermore, I already mentioned a more significant difference: on all the dynamic accounts, the *truth* of the second sentence of a reverse Sobel sequence depends on what possibilities have been raised. The analogous claim about the zebra sequence would be that the truth of 'that animal was born with stripes' depends on whether someone has raised the possibility that the designated animal is a cleverly disguised mule. On my account of reverse Sobel sequences, what possibilities have been raised need not affect whether a Sobel sequence sentence is true, but only what a speaker must do to responsibly utter the sentence. This is a key difference between the dynamic accounts and my own.

## 1.4 Arguments for my analysis

One reason to prefer my analysis to a dynamic semantics is that it is an independently motivated, more general theory. There must be some explanation for

14. I do not endorse this theory. Presuppositions are essentially marked by the way they project through some environments and not others, yet (3a) and (5a) make the relevant possibilities salient regardless of what linguistic environments they are in. But this presupposition-based theory has a lot in common with my proposal, so it is instructive to contrast this theory with the dynamic accounts, to highlight differences between the dynamic accounts on the one hand, and the presupposition-based theory and my proposal on the other.

why the zebra sequence (5) is bad. Once we have developed (EI) to account for (5), we get an explanation for reverse Sobel sequences for free. Gillies and von Fintel, on the other hand, posit semantic rules specifically to account for the infelicity of sequences of counterfactuals. The rules are part of the lexicon. (EI) explains the same data by appealing to general, independently plausible facts about conversation and reasoning. So my analysis shares a general virtue of pragmatic theories: it explains more, using less.

Another reason to prefer my analysis is that it more accurately predicts our judgments about a wide range of data. In order to adjudicate between my analysis and the dynamic semantics, we should try to find cases where the two theories make different predictions. In fact, we can find such cases. I said in §3 that certain sequences of counterfactuals are *generally* infelicitous, because the conditions in (EI) are generally met when they are uttered. But there are exceptions to these generalities. In exceptional cases, some conditions in (EI) will fail. My analysis and the dynamic semantics yield different predictions about these cases.

For example, my analysis naturally explains our intuitions about cases in which condition (iv) of (EI) fails. Remember that (3b) is generally infelicitous because speakers of (3b) are generally asserting propositions incompatible with salient possibilities that they cannot rule out. For instance, a speaker who utters (3b) after (3a) generally cannot rule out the possibility that Sophie might have been stuck behind a tall person if she'd gone to the parade. But these generalizations about speaker ignorance do not apply to every reverse Sobel sequence scenario. In some cases, a speaker who utters a reverse Sobel sequence may have some independent reason to utter the first sentence, a reason that would be in play even if she could rule out the trouble-making possibility made salient by that sentence. She may then utter the first sentence despite being able to rule out that possibility. In this kind of case, condition (iv) of (EI) may not hold for the second sentence in the sequence. My analysis predicts that the second sentence of this sort of reverse Sobel sequence will not be infelicitous in the way that a typical reverse Sobel sequence is.

And indeed, this is just what we find. Suppose John and Mary are our mutual friends. John was going to ask Mary to marry him, but chickened out at the last minute. I know Mary much better than you do, and you ask me

whether Mary might have said yes if John had proposed. I tell you that I swore to Mary that I would never actually tell anyone that information, which means that strictly speaking, I cannot answer your question. But I say that I will go so far as to tell you two facts:

(6a) If John had proposed to Mary and she had said yes, he would have been really happy.

(6b) But if John had proposed, he would have been really unhappy.

In this reverse Sobel sequence scenario, it is okay to utter (6b) after (6a). Here is why: I still have a reason to utter (6a), even if I can rule out the possibility that Mary might have married John if he had asked her. I may utter (6a) and (6b) precisely in order to get you to rule out that possibility, without breaking my promise to Mary. In this kind of case, condition (iv) does not hold for (6b), and so (EI) does not entail that my utterance of (6b) is irresponsible.

Just the same thing can happen when you ask me for two independent pieces of information. Suppose you want to know about whether the act of proposing would have led to John being happy, and you also want to know whether Mary really would be a good partner for John. So you ask me not only whether John would have been happy if he had proposed, but also whether he would have been happy if he had successfully proposed. In this scenario, even if I can rule out the possibility that Mary might have said yes if John had proposed, I still have an independent reason to utter (6a), namely to answer your second request for information. In this kind of scenario, condition (iv) does not hold for (6b), so uttering (6b) is not irresponsible. Hence my account correctly predicts that (6b) after (6a) will not be infelicitous in the way that (3b) is generally infelicitous after (3a).

Gillies and von Fintel have trouble predicting our judgments about (6). On their theory, we cannot truly utter the second sentence of a reverse Sobel sequence after accommodating the demands of the first sentence. (6a) expands the domain over which counterfactuals quantify, so that it includes some worlds in which John proposes to Mary and she says yes. Once this happens, there is no semantic mechanism for shrinking the domain of the counterfactual. So after (6a), no utterance of (6b) should be true. This simply contradicts our intuitions about (6b) as uttered in the context described above.

Condition (iv) of (EI) may also fail when it is a stable feature of the common ground that potentially trouble-making possibilities do not obtain. For instance, even in a philosophy classroom, we are used to ruling out the possibility that kangaroos might have had crutches if they had lacked tails. So even in a philosophy classroom, the following sequence may be felicitous:

(7a) If kangaroos had lacked tails but had crutches, they would have had no trouble staying upright.

(7b) But if kangaroos had lacked tails, they would have toppled over.

Here again, a speaker who utters (7a) generally has some independent reason to utter this sentence, despite being able to rule out the possibility that kangaroos might have had crutches if they had lacked tails. In this kind of scenario, condition (iv) does not hold for (7b), and again, my account correctly predicts that the second sentence of a reverse Sobel sequence is not infelicitous in the way that (3b) generally is.

My analysis also naturally explains our intuitions about cases in which condition (iii) of (EI) fails. Consider the following reverse Sobel sequence, due to John Hawthorne:

(8a) If Sophie had gone to the parade and been shorter than she actually is, she would not have seen Pedro.

(8b) But if Sophie had gone to the parade, she would have seen Pedro.

It is easy and natural to raise the possibility that if Sophie had gone to the parade, she might have been stuck behind a tall person. It is less natural to raise the possibility that if she had gone to the parade, she might have been shorter than she actually is. Of course, it is *possible* to raise this possibility. If we have just been talking about whether parade vendors would profit from selling height-affecting drugs at large events, then it will be easier to raise the possibility that Sophie might have been shorter if she had gone to the parade. I take it that (8) is generally infelicitous in contexts like these. But in other contexts, we may willingly overlook worlds in which shorter counterparts of Sophie attend the parade. In these contexts, uttering (8a) does not suffice to raise the possibility that she might have been shorter if she had gone to the

parade. I take it that (8) is felicitous in these contexts. Gillies and von Fintel again have trouble predicting the data. On their theory, no utterance of (8b) should be true once (8a) expands the domain that counterfactuals quantify over.

Having seen two ways in which (EI) conditions can fail, we can now see the exact nature of the data to be explained. Our aim is not to explain why some sequences of counterfactuals are infelicitous and some are okay. Our aim is to explain why each sequence of counterfactuals is infelicitous as uttered in certain contexts, and okay in others. Even our original reverse Sobel sequence (3) can be felicitous. For instance, suppose you belong to a mafia organized to manipulate the exact movements of every tall person who attends a parade. If I ask you whether your mafia is conspiring to corner Sophie, you could still have a reason to tell me (3a), even if you can rule out the possibility that she might have been stuck behind a tall person if she had gone to the parade. My account correctly predicts that the typical infelicity of (3b) will not be present in these contexts.

Our judgments about reverse Sobel sequences are further complicated by the fact that speakers can signal whether (EI) conditions hold. For instance, simply in uttering the second sentence of a reverse Sobel sequence, a speaker may signal that she does not satisfy condition (iv) of (EI). Sending this signal is especially easy when her audience is not sure what she knows. Moreover, a speaker may strengthen this signal in a number of ways, e.g. by speaking assertively, adopting a condescending tone, or responsibly acknowledging that her assertion has contentious consequences. For example, the following sequence may end up sounding perfectly fine:

(9a) If Sophie had gone to the parade and been stuck behind a tall person, she would not have seen Pedro.

(9b) But hey, listen up—I am telling you: if she had gone, she would have seen him.

Speakers may also signal that they wish to ignore certain salient possibilities for purposes of conversation. Deliberately ignoring possibilities is sometimes signalled by a tone of impatience. It is also easier to deliberately ignore possi-

bilities which are taken to be improbable:

(10a) If Sophie had gone to the parade and been stuck behind a tree, she would not have seen Pedro.

(10b) Oh, *come on*—if she'd gone, she would have seen Pedro.

Deliberately ignoring possibilities is a way of ruling them out of consideration. Of course, it may be in some sense irresponsible to deliberately ignore possibilities. (EI) concerns only one kind of epistemic irresponsibility: the kind that comes when a speaker neglects salient live possibilities. If you rule out salient possibilities by deliberately ignoring them, then as far as (EI) is concerned, you are not irresponsibly neglecting those possibilities. That is why (10b) is not generally infelicitous in the way (3b) is generally infelicitous.

My analysis not only explains our intuitions about felicitous reverse Sobel sequences; it also explains our intuitions about infelicitous counterfactuals in other linguistic contexts. Consider the following sequence:

(11a) Sometimes tall people go to parades and keep anyone who is behind them from seeing much of the parade.

(11b) #But if Sophie had gone to the parade, she would have seen Pedro.

(11a) is not a counterfactual. But it nevertheless raises the possibility that if Sophie had gone to the parade, she might have been stuck behind a tall person. My analysis predicts that (11b) is therefore infelicitous. Gillies and von Fintel do not predict this. Since (11a) is not a counterfactual, or even a modal sentence, it does not prompt any expansion of the domain over which counterfactuals quantify. So the dynamic semantic theory predicts that (11b) should sound as good as the second sentence of a Sobel sequence.

None of these sequences is quite a *counterexample* to the dynamic semantics. Strictly speaking, Gillies and von Fintel could alter or add to their semantics to accommodate our intuitions about these sequences. They could say that in the felicitous sequences, some mechanism prevents the domain that counterfactuals quantify over from expanding as it usually does when the earlier counterfactuals are uttered. Or they could say that some mechanism shrinks the domain before the final felicitous counterfactual is uttered. For instance, Gillies and

von Fintel could say that the domain of the counterfactual shrinks when (6b) is uttered, so that it no longer includes worlds in which Mary accepts John's proposal. Similarly, Gillies and von Fintel could say that some mechanism expands the domain of the counterfactual when sentences like (11a) are uttered, so that sentences like (11b) are infelicitous afterwards.

But the tables have turned: now the standard semantics—together with my pragmatic analysis—is the stronger theory, yielding systematic predictions about when counterfactuals are infelicitous. The dynamic theories in VON FINTEL 2001 and GILLIES 2007 aim to explain the infelicity of some counterfactuals as a semantic effect of uttering others. I have argued that the data do not support the dynamic theory: we find felicitous counterfactuals after purportedly troublesome counterfactuals have been uttered, and infelicitous counterfactuals in the absence of troublesome counterfactuals. Meanwhile, the standard semantics does not need to postulate any ad hoc shifting of the contextually determined similarity ordering to explain our intuitions about reverse Sobel sequences, or other sequences of counterfactuals I have discussed. The data support a pragmatic theory of these sequences: counterfactuals are infelicitous when the conditions of (EI) hold, and can be felicitous when those conditions fail. In addition to being more general and independently motivated, the pragmatic account in §3 explains our intuitions about a wide range of uses of counterfactuals.

## 1.5 'Might' counterfactuals

Our project is far from over. I will end with a few remarks about another potential application of the pragmatic approach: infelicitous sequences containing 'might' counterfactuals. Playing devil's advocate for the semantic approach, I will give one reason to think my §3 account does not fully explain why these sequences are infelicitous. But remaining optimistic about a pragmatic approach, I will state some desiderata for an analysis of 'might' sequences, and argue that the stated desiderata rule against some popular semantic accounts of their infelicity.

Consider the following sequence:

(12a) If Sophie had gone to the parade, she might have missed Pedro.

(12b) #But if Sophie had gone to the parade, she would have seen Pedro.

How should we explain the infelicity of (12b) after (12a)? Recall that Lewis accepts that 'might' and 'would' counterfactuals are duals. So Lewis could say that (12b) sounds bad because it is incompatible with (12a). However, we need not limit ourselves to a semantic explanation of the infelicity of (12). Given the success of the pragmatic analysis so far, we might expect an alternative, pragmatic explanation of this infelicity.

However, extending the pragmatic approach to (12) is not straightforward. There is reason to think that (EI) does not fully explain why (12) is infelicitous. (12b) sounds bad when asserted after (12a). But in addition, (12) sounds bad in the context of a supposition:

(13) #Suppose that if Sophie had gone to the parade, she might have missed Pedro, but that if she had gone to the parade, she would have seen Pedro.

Something must explain why (13) is infelicitous. And (EI) alone cannot explain it. (EI) says that it is epistemically irresponsible to *express* a proposition incompatible with a salient live possibility. But that is not something that I do when I ask you to *suppose* that if Sophie had gone to the parade, she would have seen Pedro.

To see this point another way, remember that the conditions of (EI) are generally met when a speaker utters (5b):

(5a) Cleverly disguised mules are not born with stripes.

(5b) #But that animal was born with stripes.

But it is perfectly okay to utter (5b) after (5a) in the context of a supposition:

(14) Suppose that cleverly disguised mules are not born with stripes, but that that animal was born with stripes.

The conditions of (EI) fail for (5), but it is okay to suppose (5). So it cannot be bad to suppose (12) only because the conditions of (EI) fail for (12). Something extra is wrong with (12), something that explains why it is infelicitous even in the context of a supposition.

I think that (12) is not fundamentally different from a traditional reverse Sobel sequence: (12) is infelicitous for pragmatic reasons. Defending a pragmatic account would involve defending claims about the semantics of 'might' and 'would' counterfactuals, and about the behavior of various modals in the context of suppositions. I will not thoroughly defend a pragmatic account of (12) here. However, I will state two desiderata which make trouble for some semantic accounts of (12).

Desideratum one: an account of (12) should recognize similarities between (12) and (15):

(15a) Sophie might not see Pedro.

(15b) #But Sophie will see Pedro.

Note that like (12), (15) continues to be infelicitous in the context of a supposition:

(16) #Suppose that Sophie might not see Pedro, but that she will see Pedro.

Ideally, we would give similar explanations for these similar data involving 'would' counterfactuals and future contingents. The same goes for conditionals embedding future contingents, such as:

(17a) If Sophie goes to the parade, she might not see Pedro.

(17b) #But if she goes to the parade, she will see Pedro.

Some theorists give a similar semantics for pairs of conditionals such as:

(3b) But if she had gone to the parade, she would have seen Pedro.

(17b) But if she goes to the parade, she will see Pedro.

If we accept a unified theory of "had-would" and "does-will" conditionals, there is even more pressure to find a unified explanation of the infelicity of (12) and (17). Ultimately, these sequences may not be infelicitous for exactly the same reason. But minimally, an account of (12) should accommodate judgments about similar sequences containing future contingents.

Desideratum two: an account of (12) should explain the embedding behav-

ior of (12b). For instance, (12b) is not felicitous after (12a), but neither is its negation:

(12a) If Sophie had gone to the parade, she might have missed Pedro.

(12b) #But if Sophie had gone to the parade, she would have seen Pedro.

(12c) Hey, look, you can't say that, because you don't know whether she would have seen him if she'd gone. / #Hey, look, you can't say that, because it's just false that she would have seen him if she'd gone.

It is not felicitous to deny (12b) when it is embedded in a question:

(18a) If Sophie had gone to the parade, she might have missed Pedro.

(18b) So would she have seen him if she'd gone?

(18c) I don't know for sure. / #No.

It is okay to assign (12b) a high subjective probability:

(19a) If Sophie had gone to the parade, she might have missed Pedro.

(19b) But I suspect that she would have seen him, if she'd gone.

In these sequences, there is a striking contrast between the embedding behavior of (12b) and sentences that are generally agreed to be false, such as (20):

(20) If Sophie were to go to the parade, she would definitely see Pedro.

Unlike (12b), the embedding behavior of (20) confirms that it is false:

(21a) If Sophie had gone to the parade, she might have missed Pedro.

(21b) #But if Sophie had gone to the parade, she would have definitely seen Pedro.

(21c) #Hey, look, you can't say that, because you don't know whether she would definitely have seen him if she'd gone. / Hey, look, you can't say that, because it's just false that she would definitely have seen him if she'd gone.

(22a) If Sophie had gone to the parade, she might have missed Pedro.

(22b) So would she definitely have seen him if she'd gone?

(22c) #I don't know for sure. / No.

(23a) If Sophie had gone to the parade, she might have missed Pedro.

(23b) #But I suspect she would definitely have seen him, if she'd gone.

The embedding behavior of (12b) suggests that (12b) is not straightforwardly false: it is more natural to negate and deny false utterances, and less natural to assign them high probability. Explanations of why (12b) is bad should at least accommodate these data, if not predict them.

These desiderata rule against an increasingly popular hypothesis about counterfactuals, recently defended in HÁJEK 2007. Hájek claims that 'might' counterfactuals like (12a) are almost always true:

(12a) If Sophie had gone to the parade, she might have missed Pedro.

(12b) #But if she had gone to the parade, she would have seen Pedro.

Hájek observes that (12) sounds contradictory. He concludes that (12a) and (12b) are contraries, and that (12b) is therefore false. One may repeat this argument for most counterfactuals. On these grounds, Hájek concludes that most 'would' counterfactuals are false.

In stating desiderata for a theory of (12), I have raised two worries for Hájek's argument. My first worry: one could use the same strategy to argue that most future contingents are false, on the grounds that sequences like (15) sound contradictory. But this would be an unwelcome conclusion. Some think that past utterances of future contingents had or have indeterminate truth values. But it is hard to accept that utterances of future contingents are automatically *false*. For instance, most theorists strongly resist saying my utterance yesterday of 'I will be alive tomorrow' is or was false, when I am in fact alive today.

My second worry: our judgments about (12c), (18), and (19) suggest that (12b) is not false. I think that Hájek might respond to this worry by saying that our judgments about (12c), (18), and (19) do not accurately signal whether (12b) is false. Hájek says that our practice of uttering counterfactuals such as (12b) is

"legitimated" by the existence of "nearby" true counterfactuals such as (24):

(24) If Sophie had gone to the parade, she would very probably have seen Pedro.

Hájek says that since (24) is true, and closely related to (12b), we may legitimately assert the latter:

> There are true counterfactuals closely related to the ones we assert that support our practice, at least when the prevailing standards for asserting counterfactuals are somewhat forgiving, as they typically are on the street. So we can legitimately assert various counterfactuals. Still, most of them remain false. (52)

Hájek could go on to say that since (24) is closely related to (12b), we may legitimately judge embedded occurrences of (12b) as if they were occurrences of (24). Moreover, we do in fact judge embedded occurrences of (12b) in this way. So our judgments about (12c), (18), and (19) reflect whether (24) is false, not whether (12b) is false.

To respond: we do not in fact judge embedded occurrences of (12b) as if they were occurrences of (24). For instance, compare the following sequences, as uttered by speakers who know that Sophie is an extremely tall and aggressive Pedro fan:

(19a) If Sophie had gone to the parade, she might have missed Pedro.

(19b) So would she have seen him if she'd gone?

(19c) I don't know for sure.

(25a) If Sophie had gone to the parade, she might have missed Pedro.

(25b) So would she have very probably seen him if she'd gone?

(25c) #I don't know for sure.

It is more natural to attribute knowledge of propositions expressed by counterfactuals such as (24), and harder to attribute knowledge of propositions expressed by counterfactuals such as (12b). Our judgments about (19) and (25) distinguish (12b) from (24). So it is reasonable to assume that our judgments about (12c), (18), and (19) suggest that it is (12b) itself, not just (24), that is not

false. This leaves open several explanations of why (12b) is infelicitous. But it does tell against semantic explanations of the sort that Hájek gives.

## 1.6 Constraining credences in counterfactuals

So far I have defended both STALNAKER 1968 and LEWIS 1973a, without dwelling on differences between their accounts. But though these accounts are similar in spirit, they famously differ over what truth value to assign to counterfactuals such as:

(26) If I had flipped this coin, it would have landed heads.

It is often taken for granted that in normal contexts, no world counts as the unique closest world in which the speaker flips a fair coin. Instead there are equally close worlds in which the coin is flipped and lands heads or lands tails. Lewis concludes from this that in such contexts, (26) is false. Stalnaker concludes that (26) is indeterminate, neither true nor false.

One might expect these different conclusions to count in favor of one semantics or the other. In particular, one might expect our judgments about embedded occurrences of (26) to give us evidence about whether (26) is false or indeterminate. Recently, EDGINGTON 2008 has argued that this is not the case. Edgington claims that our judgments about embedded occurrences of (26) make trouble for *both* semantic theories. Edgington argues that since traditional truth-conditional theories are unable to account for these embedding data, we should adopt a non-truth-conditional theory of counterfactuals. In the rest of this chapter, I introduce Edgington's concern and defend truth-conditional theories of counterfactuals against her criticism. Though embedding data tell against the particular semantic theory defended in HÁJEK 2007, they lend general support to semantic theories assigning truth conditions to counterfactuals.

## 1.7 Judgments about embedded counterfactuals

Recall that on the semantics for counterfactuals given in LEWIS 1973a, counterfactuals are context-sensitive. In particular, the context in which a counterfactual is uttered contributes a similarity ordering on worlds to its truth conditions.

Given that $p$ is possible, Lewis says that 'if $p$, would $q$' expresses a proposition that is non-vacuously true at a world just in case some $p$-and-$q$ world is closer to that world than any $p$-and-not-$q$ world, according to the contextually determined similarity ordering. For example, Lewis says that (26) is true just in case some world where the coin is flipped and lands heads is closer than any world where the coin is flipped and lands tails. Otherwise, (26) is false.

In Lewis 1973b, Lewis talks about worlds being tied for closest according to the contextually determined similarity ordering. He discusses the following counterfactuals from Quine 1950:

(27) If Bizet and Verdi had been compatriots, Bizet would have been Italian.

(28) If Bizet and Verdi had been compatriots, Verdi would have been French.

Lewis says that when (27) and (28) are uttered, "the case may be perfectly balanced between respects of comparison that favor the [world where both composers are French] and respects that favor the [world where both are Italian]" (60). And when the case is perfectly balanced, Lewis says that both (27) and (28) are false. Given these remarks, it is natural to think that when a speaker utters (26) about a fair coin, Lewis will say that the case is similarly balanced, and that (26) is false.

The same conclusion is suggested by Lewis's theory of 'might' counterfactuals. Intuitively, the following is true in normal contexts:

(29) If I had flipped the coin, it might have landed tails.

For Lewis, 'might' and 'would' counterfactuals are duals. So since (29) is true, Lewis will say that (26) is false. Moreover, if we know that the coin in question is fair, then we know that (29) is true. Insofar as the duality of 'might' and 'would' is apparent to us, Lewis will say that we know that (26) is false.

Suppose that Lewis is right, and we do know that (26) is false. Then we should not only be unwilling to assert (26); we should be willing to deny it. For instance, we should assent to the following:

(30) It's not the case that if I had flipped the coin, it would have landed heads.

(31) It's not the case that if Bizet and Verdi had been compatriots, Bizet would have been Italian.

(32) It's not the case that if Bizet and Verdi had been compatriots, Verdi would have been French.

But on a natural reading of (30), it sounds as bad as (26). And (31) and (32) are no better. STALNAKER 1978 observes that "most speakers would be as hesitant to deny as to affirm either of the conditionals [(31) or (32)]" (92).

One might respond on Lewis's behalf by insisting that counterfactuals like (30)-(32) are indeed assertable. For example, HÁJEK 2007 claims that most 'would' counterfactuals are false, on the grounds that one can often find a true 'might' counterfactual contrary to any given 'would' counterfactual. Hájek suggests that embedding data are no problem for his theory, because it is perfectly fine to assert counterfactuals such as:

(33) It's not the case that if I had flipped the coin, it WOULD have landed heads.

So Hájek is not motivated to give a semantics for counterfactuals that predicts that (30)-(32) are unassertable.

I discussed some shortcomings of Hájek's theory of counterfactuals in §5. Here is another reason to think that Hájek overlooks at least some readings of counterfactuals like (30): by capitalizing 'would' when giving examples of unassertable counterfactuals, Hájek prompts readings of the counterfactuals where 'would' is focused. But focusing constituents may change the semantic values of counterfactuals. By concentrating on focused readings of counterfactuals, Hájek may overlook other available readings.

If a speaker does not focus 'would' in (30), then there is at least one reading on which (30) is unassertable. Just as with the counterfactuals discussed in §5, one can isolate this reading of (30) by imagining that it answers a question:

(34a) Hey, just make a guess: if you had flipped the coin, would it have landed heads? Or would it have landed tails?

(34b) #Well, it's not the case that it would have landed heads.

If it is common ground that the coin in question is fair, (34b) sounds bad in response to (34a). Insofar as (34b) sounds acceptable, it sounds as if the speaker of (34b) knows that neither side of the coin is marked heads. This is the reading of (30) that I am interested in accounting for. On this reading, (30) is unassertable if it is common ground that the coin in question is fair. Far from asserting or assuming (30), it is felicitous to ask someone to guess whether (30) holds. On at least one reading, (30)-(32) should be unassertable.

These intuitions about (30)-(32) appear to pose a problem for Lewis's theory of counterfactuals. Our reluctance to assert (30)-(32) appears to be at odds with the Lewisian verdict that we know that (26)-(28) are false and that (30)-(32) are true. In discussing his rejection of the conditional law of excluded middle, even Lewis suggests that his semantics does not respect the "offhand opinion of any ordinary language speaker," and that *ceteris paribus*, one should aim to respect such judgments (Lewis 1973a, 80). Edgington 2008 argues that counterfactuals like (30) pose a serious problem for Lewis: namely, in plenty of cases, we are unwilling to deny a counterfactual that Lewis says we know is false (13).

Stalnaker 1978 accommodates more of our ordinary language judgments about embedded counterfactuals such as (30)-(32). Like Lewis, Stalnaker says that counterfactuals are context-sensitive. In particular, the context in which a counterfactual is uttered contributes a selection function to its truth conditions. Roughly speaking, the selection function maps any proposition and world to the unique closest world in which the given proposition is true.

Stalnaker says that 'if $p$, would $q$' is true at a world just in case this unique closest $p$-world is a $q$-world. Stalnaker says that sometimes "the selection functions that are actually used in making and interpreting counterfactual conditional statements correspond to orderings of possible worlds that admit ties and incomparabilities" (90). In such cases, counterfactuals are neither true nor false. So for Stalnaker, counterfactuals such as (27) and (28) are not false, but indeterminate in truth value. Hence Stalnaker can account for our reluctance to affirm (30)-(32): just like their negations, these counterfactuals are neither true nor false.

However, other judgments about embedded counterfactuals seem to be just as challenging for Stalnaker as for Lewis. For example, it sounds okay to won-

der whether a counterfactual is true:

(35) I wonder whether the coin would have landed heads if I had flipped it.

One can also embed counterfactuals in credence ascriptions:

(36) I think it is exactly .5 likely that if I had flipped the coin, it would have landed heads.

Of course, these judgments seem to make trouble for Lewis. If we know a claim is false, it is hard to see how we could assign it .5 credence, or wonder whether it is true. But these judgments also seem to make trouble for Stalnaker. It is hard to see how we could assign a claim exactly .5 credence, or wonder whether it is true, if we know that the claim is neither true nor false. So our judgments about (35) and (36) seem to be at odds with a supervaluationist semantics on which (26) is indeterminate in truth value.

One way to appreciate this apparent difficulty for Stalnaker is to compare his predictions about counterfactuals with supervaluationist predictions about vague sentences. On a supervaluationist semantics, the following sentence is indeterminate in truth value when John is a borderline case of baldness:

(37) John is bald.

The supervaluationist treats vague borderline sentences like Stalnaker treats ordinary counterfactuals. If Stalnaker gives the correct semantics for counterfactuals, one would expect our judgments about (26) to resemble a supervaluationist's judgments about (37).

But when it comes to embedded occurrences, we do not in fact judge (26) as a supervaluationist judges (37). Faced with a clear borderline case of baldness, a supervaluationist will not assert the analog of (35) or (36):

(38) #I wonder whether John is bald.

(39) #I think it is exactly .5 likely that John is bald.

Insofar as you are a supervaluationist, you think that there is no determinate fact of the matter about whether John is bald. In this frame of mind, it sounds extremely unnatural to wonder about whether a borderline case of baldness is

bald, or to assign .5 credence to such a claim. The ascriptions in (35)-(36) and (38)-(39) sound appropriate only insofar as they successfully invoke epistemicist assumptions: the claim that there is a fact of the matter that the speaker is wondering and forming opinions about. Our judgments about (35) and (36) suggest that we naturally assume there is a fact of the matter about counterfactuals. Given these judgments, it may seem hard for a supervaluationist theory to account for all our ordinary uses of counterfactuals.

EDGINGTON 2008 focuses on ordinary language judgments about assignments of credence to counterfactuals. She argues that we should reject the semantic theories in LEWIS 1973a and STALNAKER 1968, on the grounds that they fail to accommodate many such judgments. In particular, Lewis predicts that we assign high credence to many counterfactuals that we know to be false. Stalnaker faces a similar charge: "vast numbers of subjunctive conditionals just get the verdict 'indeterminate' and this is not very helpful" (18). Edgington concludes that our semantics should not assign truth conditions to counterfactuals: "when we consider the uncertain judgments we express as subjunctive conditionals, the case is just as strong as it is for indicatives, that they do not express propositions" (3).

Edgington discusses credence ascriptions, but we have seen many ways in which our judgments about counterfactuals indicate that we do not talk as if these counterfactuals are false or indeterminate in truth value. This holds when we ask and answer questions embedding counterfactuals, when we are reluctant to assert their negations, and when we embed counterfactuals in ascriptions of attitudes like wondering. Edgington worries that our ordinary language judgments cannot be accommodated by a semantic theory that assigns truth conditions to counterfactuals. On these grounds, she concludes that we should adopt a non-truth-conditional semantics for counterfactuals.

### 1.8 Truth-conditional accounts of embedding data

I agree with Edgington that our judgments about embedded counterfactuals indicate that we do not talk as if counterfactuals about chance events are false or indeterminate in truth value. Rather, we wonder and form opinions about counterfactuals. In other words, we talk as if such sentences could turn out to

be true. Unlike Edgington, I think a truth-conditional theory of counterfactuals can accommodate these data.

The easiest way to accommodate the data is to allow that counterfactuals such as (26)-(28) can indeed be true in normal contexts. One can give such a truth-conditional theory by accepting the limit and uniqueness assumptions, and by admitting that context can contribute a single similarity ordering to the truth conditions of a counterfactual. For instance, suppose the truth conditions of counterfactuals like (26)-(28) are those given by Stalnaker or Lewis, and also suppose that in normal contexts, there is a unique closest world in which the antecedents of such counterfactuals are satisfied. Then each ordinary counterfactual may be true or false, depending on the nature of the closest world where the antecedent is satisfied. In normal contexts, an ordinary speaker may not know much about which world is the closest antecedent world. So she may not know which truth value the counterfactual has.

This theory easily accounts for our judgments about embedded counterfactuals. Since you do not know whether the closest world in which I flip the coin is one where the coin lands heads, you are not in a position to assert (30). It makes sense for you to wonder about the nature of the closest world where I flip the coin, as in (35). And we can explain why you may appropriately assign a particular credence value to a counterfactual proposition, as in (36). For example, you may be uncertain about whether the closest world in which I flip the coin is one where it lands heads or one where it lands tails. Perhaps you give .5 credence to each hypothesis. In that case, you may appropriately assign .5 credence to the proposition that if I had flipped the coin, it would have landed heads. Our assignments of credence to counterfactuals make sense: these assignments reflect our opinions in the face of uncertainty about where our world is located with respect to other possible worlds.

Endorsing the claim that there is a closest antecedent world allows us to account for other patterns of ordinary language use as well. For instance, the following question carries an existence presupposition:

(40) If I had tossed this coin, which side would have landed up?

In particular, (40) presupposes that there is a side of the coin such that if I had

tossed the coin, that side would have landed up.[15] So in using (40), we presuppose that there is a fact of the matter about (26). Moreover, we sometimes use phrases like "the way things would have been if I had flipped the coin." In addition to our judgments about counterfactuals under negation and attitude operators, constructions like these are more evidence that we talk as if context contributes a closest antecedent world to the truth conditions of a counterfactual.

There are some costs to accepting this theory of counterfactual talk. In particular, the claim that context determines a closest antecedent world requires two kinds of determinacy: determinacy in language and determinacy in the world. The features of a context must be detailed enough to determine a similarity ordering that singles out a particular world as closer than any other. If context does determine such a similarity ordering, we must adopt an epistemicist theory of how this happens. Facts about exactly which world is the closest antecedent world would be inaccessible to speakers, just as epistemicists think facts about the exact extension of a vague predicate are inaccessible.

Both Lewis and Stalnaker reject determinacy in language. LEWIS 1973b says that the claim that there is a closest antecedent world is "thoroughly implausible" (60), and STALNAKER 1978 says it is "grossly implausible" (89). Lewis says it is implausible that we manage to single out a closest antecedent world "despite the infinite number and variety of worlds" (60). Stalnaker says "it is unrealistic to assume that our conceptual resources are capable of well ordering the possible worlds" (90). These arguments seem quick in light of more recent defenses of epistemicism given by WILLIAMSON 1994 and others. Many vague terms have an infinite number and variety of possible denotations, but epistemicism about vagueness is nonetheless defensible. It is unrealistic to assume that our conceptual resources are capable of distinguishing between all of these possible denotations, but WILLIAMSON 1994 argues that facts about our linguistic community, beyond facts our conceptual resources, help fix the denotations of vague terms.

The assumption that there is a closest antecedent world also depends on determinacy in the world. Suppose there are genuinely chancy events. One

15. See POSTAL 1971 and KARTUNNEN & PETERS 1976 for discussion of existence presuppositions carried by wh-questions.

might reason as follows: if counterfactuals are true, then there must be facts grounding their truth. For instance, if 'if I had flipped the coin, it would have landed heads' is true, there must be some fact about the actual world that makes it true. Any such fact would fail to supervene on physical facts, and would therefore be metaphysically suspect. So any metaphysically scrupulous person should hold that counterfactuals lack truth values. Fans of the assumption that there is a closest antecedent world must reject this reasoning and admit brute metaphysical facts into their ontology. Of course, we do not have to single out exactly one antecedent world: as long as all closest antecedent worlds are uniform with respect to whether the consequent holds, Lewis and Stalnaker will predict that a counterfactual has a determinate truth value. But in the case of counterfactuals with consequents about chancy events, accepting homogeneity seems as problematic as accepting the uniqueness assumption. If uniqueness fails, one may still wonder how context and the world could ever be rich enough to select a set of closest antecedent worlds where only one of a number of chance outcomes occurs.

In light of this debate, the key point to recognize is that the truth-conditional theorist does not need to argue for homogeneity. She can account for our judgments about embedded counterfactuals, however the debate about homogeneity goes. If homogeneity holds, she may say that ordinary speakers recognize that it does. If homogeneity does not hold, she may say ordinary speakers nevertheless talk as if it does.

Stalnaker endorses the latter strategy. For instance, consider the counterfactual question:

(41) If President Kennedy had not been assassinated in 1963, would the United States have avoided the Vietnam War debacle?

Stalnaker makes just a few brief remarks about how we typically respond to such questions:

> Even when we recognize that such a question really has no answer, we continue to talk and think as if there were an answer that we cannot know. This is, I think, because we tend to think of the counterfactual situations determined by suppositions as being as complete and determinate as our own actual world. (102)

In other words, Stalnaker suggests that even though there is no closest antecedent world, we talk as if there is. One might argue that in this respect, our talk about counterfactuals is like our talk about personal identity.[16] In ordinary language discussions about particular cases, we act as if there are determinate, non-supervening facts about personal persistence, facts that we do not know and may never discover. But on reflection, those who are skeptical of brute, non-supervening facts may reject this talk as misguided. It may be especially efficient or productive for ordinary speakers to talk as if there is a fact of the matter in some cases of genuine indeterminacy. In such cases, a reflective theorist can endorse a fictionalism about the relevant discourse.

Several linguists have argued for a formal version of the claim that we talk as if homogeneity holds: that counterfactuals semantically presuppose homogeneity. Some have suggested that 'if $p$, would $q$' is best analyzed as a generic bare plural construction, so that the correct analysis of (42a) is given by (42b):

(42a) If it were the case that $p$, it would be the case that $q$.

(42b) $P$ worlds are $q$ worlds.

For instance, VON FINTEL 1997 argues that (42b) is the correct analysis of (42a). Given one plausible analysis of generic sentences, context fixes a set of $p$ worlds relevant to the truth of (42b). In the case of counterfactuals, the relevant $p$ worlds are just those closest to ours. Once we accept this analysis, we can infer facts about what (42a) presupposes. FODOR 1970 argues that generic bare plurals carry an "all-or-none" presupposition, so that (42b) presupposes that either all $p$ worlds are $q$ worlds, or none of them are.[17] Since (42b) is the correct analysis of (42a), von Fintel concludes that (42a) carries the same presupposition of homogeneity.

Other linguists have suggested that 'if $p$, would $q$' is best analyzed as a definite plural construction, so that (42c) is the correct analysis of (42a):

(42c) The $p$ worlds are $q$ worlds.

FODOR 1970 argues that definite plurals carry the same "all-or-none" presuppo-

16. Thanks to Ned Hall for the comparison.

17. See LÖBNER 1987 for more recent work on homogeneity presuppositions.

sition as generic bare plurals. So by saying that the analysis of (42a) is given by (42c), we can again conclude that (42a) presupposes that homogeneity holds.[18]

There are other ways to formalize the claim that speakers talk as if homogeneity holds. Instead of saying that counterfactuals semantically presuppose homogeneity, the truth-conditional theorist could say that speakers presuppose that homogeneity holds when they utter counterfactuals. Or she could say that speakers of counterfactuals pragmatically implicate that homogeneity holds.[19] Or she may even go so far as to say that even though homogeneity does not hold, ordinary speakers believe that it does. In this last case, she would account for ordinary judgments about counterfactuals with a simple error theory.

To sum up so far: advocates of truth-conditional theories can account for many of our ordinary judgments about counterfactuals. In fact, several accounts of our ordinary judgments are compatible with semantic theories like those given by Lewis and Stalnaker. Embedding data do not constitute an argument against truth-conditional theories of counterfactuals.

## 1.9 Grounding norms governing credence in counterfactuals

If we can account for our judgments without giving up truth-conditional semantic theories, it is reasonable to prefer this more developed and less radical approach. By saying that we talk as if homogeneity holds, we can account for how we ask and answer questions embedding counterfactuals, refuse to assert their negations, and embed counterfactuals in ascriptions of attitudes like wondering.

There is just one more complicated kind of judgment left to explain. Speakers not only ascribe specific credences to counterfactuals; they do so in a rule-governed way. For example, it is a common intuition that agents should give exactly .5 credence to the claim that a fair coin would have landed heads if it had been flipped. Theories of counterfactuals face a challenge: why do ordinary speakers endorse this norm?

The truth-conditional theories given in §2 can answer this challenge. Suppose that we talk as if homogeneity holds. Then our counterfactual credences

18. For arguments in favor of the definite plural analysis of counterfactuals, see Schein 2001 and Schlenker 2004.

19. See Krifka 1996 for an account of homogeneity involving pragmatic strengthening.

are just credences in particular propositions. Several plausible principles of epistemology and folk physics govern our credences in such propositions. These principles say how our credences in counterfactuals must relate to our credences in objective chance hypotheses. This is a significant virtue of truth-conditional theories: they allow us to explain complicated facts about what norms we take to govern our credences in counterfactuals.

For sake of simplicity, I will focus on deriving one particular norm: our judgment that when I fail to flip a fair coin, you should have .5 credence that if I had flipped the coin, it would have landed heads. One can extend arguments about this example to other cases involving counterfactuals about chance events. Let $C$ be any credence function you may rationally have before I decide not to flip the coin. Let $C'$ be any credence function you may rationally have after the coin is not flipped. The argument proceeds as follows:

| | | |
|---|---|---|
| 1. | $C(ch(H\|F) = .5) = 1$ | stipulation |
| 2. | $C(H\|F) = .5$ | 1, (SP) |
| 3. | $C((F \boxrightarrow H)\|F) = .5$ | 2, Centering |
| 4. | $C[(F \boxrightarrow (F \boxrightarrow H)) \Leftrightarrow (\overline{F} \boxrightarrow (F \boxrightarrow H))] = 1$ | folk physics |
| 5. | $C((F \boxrightarrow H)\|F) = C((F \boxrightarrow H)\|\overline{F})$ | 4, folk epistemology |
| 6. | $C((F \boxrightarrow H)\|\overline{F}) = .5$ | 3, 5 |
| 7. | $C'((F \boxrightarrow H)) = .5$ | 6, Conditionalization |

The argument starts with a simple claim: you should be certain that the conditional objective chance of the coin landing heads if flipped is .5. This is not a substantive principle, but a precise way of stipulating that you are certain that the coin in question is fair. For sake of simplicity, let us restrict our attention to cases where a fair coin has some objective chance of being flipped, and exactly half as much chance of being flipped and landing heads. If you are certain that a coin is fair, then you should be certain that these circumstances obtain.

In step two, we apply an epistemic principle similar to the Principal Principle advocated by Lewis 1980, which matches credences with beliefs about objective chances:

(PP) If you are certain of the objective chance of $p$, your credence in $p$ should equal your estimate of the objective chance of $p$.

The epistemic principle we need is the *Superintendent Principle*:

(SP) Your conditional credence in $q$ given $p$ should equal your expectation of the conditional objective chance of $q$ given $p$.

The Superintendent Principle is a generalization of the Principal Principle, governing conditional as well as unconditional credences. Several theorists have advocated principles like the Superintendent Principle.[20] I do not wish to give a detailed defense of this principle, but only suggest that it is something ordinary speakers may implicitly endorse. Given this principle, your conditional credence that the coin will land heads, given that I flip the coin, should be .5.

In step three, we apply the strong centering assumption, namely the claim that no world is as close to the actual world as the actual world itself.[21] Suppose I do decide to flip the coin. Then by the strong centering assumption, there will be a single closest world in which I flip the coin: namely, the actual world. So the coin will land heads in the closest world where I flip the coin just in case the coin actually lands heads. From this we can conclude: given that I flip the coin, your conditional credence that the coin would land heads if flipped should equal your credence that the coin does in fact land heads.

Step four states our certainty of a principle of folk physics: whether the coin would land heads if flipped is counterfactually independent of whether the coin is in fact flipped. In other words, suppose that if I had flipped the coin, then the coin would have landed heads if flipped. Then even if I hadn't flipped the coin, it still would have been the case that the coin would have landed heads if flipped. And the converse also holds.

This principle of folk physics spells out a consequence of a folk theory of counterfactual tendencies, namely that they are like dispositions. Suppose we have managed to identify a coin that would have landed heads if it had been flipped. Then it is as if the coin has a certain disposition to land heads, which is manifested just in case it is flipped. And as with many normal dispositions, we suppose that whether the coin has this disposition is independent of whether

20. For instance, SKYRMS 1978 claims that the *degree of assertability* of a subjunctive conditional such as 'if $p$ were the case, $q$ would be the case' should equal the subjective expectation of the conditional objective chance of $q$ given $p$. See also VAN FRAASSEN 1980 for arguments that your credence in $p$ should equal your expected value of the objective chance of $p$.

21. For discussion of centering assumptions, see LEWIS 1973a, p.29ff.

the manifestation conditions of the disposition obtain. In other words: whether the coin would land heads if flipped is counterfactually independent of whether the coin is in fact flipped.

In step five, we apply a principle of folk epistemology. In certain special cases, events are counterfactually independent but nevertheless evidentially linked. For example, suppose that smoking does not cause cancer, but a certain gene causes cancer and also causes people to smoke.[22] In this case, whether I smoke makes no counterfactual difference to whether I have cancer. But it makes an evidential difference to whether I think I have cancer: on learning that I smoke, I should increase my credence that I have the carcinogenic gene. Folk epistemology says that learning whether I flip a coin is not like this special case of learning whether I smoke. For instance, our folk theory of coin flipping says that there is no common cause in the world that leads me to flip the coin, and also causes the coin to have the disposition to land heads if flipped. In other words, step four says that whether the coin is flipped makes no difference to whether it would land heads, and step five allows us to draw the further conclusion that learning whether the coin is flipped should make no difference to my credence that the coin would land heads.

Several specific folk theories could underwrite the physical and epistemological principles in steps four and five. For instance, it might be that counterfactual facts are causally determined before their antecedents. If there is currently some chance that the coin would land tails if flipped, it may become determined that the coin would land heads if flipped, before it is even determined whether I will flip the coin at all.

It can be natural to implicitly assume that facts about counterfactuals are determined in this way. For example, suppose I have placed a conditional bet: conditional on my flipping the coin, I have bet that the coin will land heads. Suppose that I am still deciding whether to flip the coin at all. In such a situation, I might think to myself, "I wish I knew whether if I flipped this coin, it would come up heads or tails. Then I would know whether I should flip the coin or put it away. But as it is, I can't decide what to do." It is natural to think that in wishing to know whether the coin would land heads, I am implicitly

22. For further discussion of medical Newcomb problems, see Gibbard & Harper 1978 and Eells 1982.

assuming that there is some determined fact of the matter to be known. And if it is already determined whether the coin would land heads, then we should endorse the folk principles used above: whether I flip the coin makes no difference to whether the coin land heads, and my credence about the former should make no difference to my credence about the latter. Of course, we can endorse these folk principles without claiming that the truth of a counterfactual is always determined before the truth of its antecedent. This claim about temporal priority is simply one concrete way of grounding the folk principles mentioned above.

Step six follows directly from steps three and five. And finally, in step seven, we simply apply the principle of conditionalization: namely, that you should update by conditionalizing your credence distribution on your evidence. On learning that I am not going to flip the coin, you should conditionalize on this proposition. If your prior conditional credences are as given in (31) and you update by conditionalization, your later credence that the coin would have landed heads if I had flipped it should be .5. Hence we have reached our desired conclusion: when I fail to flip a fair coin, you should have .5 credence that if I had flipped the coin, it would have landed heads.

Just to recap: our semantic theory says that 'if I were to flip this coin, it would land heads' may express a determinate proposition, and epistemological principles say your credence in this proposition is constrained by your credences in objective chance hypotheses. Folk physics says that whether this proposition holds does not depend on whether the coin is flipped. So when I decide not to flip the coin, you get no new information about whether the coin would have landed heads, and your credence that the coin would have landed heads if flipped should remain .5. In this way, several plausible principles may ground our intuitive normative judgments about assigning credences to counterfactuals.

It is tempting to think that the inference from step three to step seven must issue from a general principle: namely, that your later credence in a past subjunctive should always equal your earlier rational credence in the corresponding future subjunctive. For example, if earlier you should have .5 credence that the coin would land heads if I were to flip it, then later you should have .5 credence that the coin would have landed heads if I had flipped it. Slote 1978 and Edg-

INGTON 2003 have argued that certain counterfactuals violate this principle.[23] For example, suppose that Edgington is about to toss a fair coin. Since the coin is fair, you should have .5 credence that if you were to bet on heads, you would win. Suppose you do not bet and the coin lands heads. Then you should be certain that if you had bet on heads, you would have won. So your later rational credence in the counterfactual proposition is much higher than your earlier rational credence.

Morgenbesser conditionals do challenge some principles linking earlier and later credences in subjunctive conditionals. But they do not challenge the main argument of this section. Given the folk physics introduced above, it is not hard to explain our intuition that your credence in a Morgenbesser conditional should change when the coin tossed by Edgington lands heads. To make the explanation vivid, suppose that whether the coin would land heads if flipped is counterfactually independent of whether you will bet that the coin will land heads. If the coin is flipped and does land heads, then you learn something: namely, that the coin would land heads if flipped. Of course, this information is relevant to whether it was also true that you would have won if you had bet on heads. In light of this new information, you may rationally change your credence in the proposition that you would have won if you had bet on heads. In other words, your evidence in the Morganbesser case includes more than the claim that you flip the coin. That is why conditionalization does not license moving from step three to step seven above. However, in a case where I decide not to flip a fair coin, you do not get information about whether my coin would have landed heads if flipped. So you do not get information about whether you would have won if you had bet on heads. So your credence in the proposition expressed by 'if I had bet on heads, I would have won' should remain .5, as should your credence in the proposition that the coin would have landed heads if it had been flipped.

Once we say that ordinary speakers talk as if homogeneity holds, we can understand why ordinary speakers endorse certain norms about events like coin tosses. The above argument from folk principles simply explains how recognizing an objective symmetry in chances should constrain our credences in various propositions, such as the proposition that the coin lands heads in

23. SLOTE 1978 attributes these conditionals to Sydney Morgenbesser.

the closest worlds where it is flipped. Once a speaker supposes that whether a counterfactual holds depends on whether a coin lands heads in the closest flip worlds, we can explain her uncertainty about the former proposition in terms of her uncertainty about the latter.

It is not clear how a non-truth-conditional theory can give just as good an explanation of norms governing our credence in counterfactuals. Edgington claims that we talk as if counterfactuals do not have truth conditions. She claims that we talk as if our credences in counterfactuals are irreducibly conditional probabilities. Just as our indicative conditional credences cannot be reduced to credences in single propositions, our counterfactual credences must be irreducible. It is not clear what principles should constrain such irreducible properties of our doxastic states. Edgington may try to adopt some analog of the derivation given above. But it is not clear how non-truth-conditional counterfactuals embed under indicative or counterfactual suppositions, or what principles link one kind of embedded counterfactual with another. The challenge remains for non-truth-conditional theories to explain how speakers arrive at their counterfactual credences, and why their opinions about objective chances should matter when they do.

Edgington could concede that speakers talk as if homogeneity holds, but still maintain that counterfactuals actually lack truth conditions. This move seems unmotivated, since accounting for embedding data was a large part of what led Edgington to reject truth-conditional accounts. But the larger problem with this move is one of theoretical economy. It is simple to model how speakers of sentences with truth conditions talk as if a particular proposition is common ground. It is harder to model how speakers systematically and productively talk as if an entire realm of discourse has truth conditions, when in fact it does not. Once we have already developed truth-conditional theories in order to explain how speakers use and evaluate embedded counterfactuals, a truth-conditional semantics for counterfactuals is the more attractive theory.

## 2 Scoring rules and epistemic compromise

Formal models of epistemic compromise have several fundamental applications. Disagreeing agents may construct a compromise of their opinions to guide their collective action, to give a collective opinion to a third party, or to determine how they should update their individual credences. Recent literature on disagreement has focused on certain questions about epistemic compromise: when you find yourself disagreeing with an epistemic peer, under what circumstances, and to what degree, should you change your credence in the disputed proposition? ELGA 2006 and CHRISTENSEN 2007 say you should compromise often and deeply; KELLY 2007 disagrees. But these authors leave open another question: what constitutes a perfect compromise of opinion?

In the disagreement literature, it is sometimes assumed that if we assign different credences $c_a$ and $c_b$ to a proposition $p$, we reach a perfect compromise by *splitting the difference* in our credences. In other words: to adopt a perfect compromise of our opinions is to assign credence $.5(c_a + c_b)$ to $p$. For instance, KELLY 2007 says that when peers assign .5 and .7 to a proposition, to adopt a compromise is to "split the difference with one's peer and believe the hypothesis to degree .6" (19).[1]

But why does .6 constitute a perfect compromise? Of course, .6 is the arithmetic mean of .5 and .7. But why must a compromise of agents' opinions always be the arithmetic mean of their prior credences? In other cases of compromise, we do not simply take it for granted that the outcome that constitutes a perfect compromise is determined by the arithmetic mean of quantities that reflect what individual agents most prefer. Suppose we are running partners, and I want to run one mile, while you want to run seven. Running four miles may not be a perfect compromise, especially if I strongly prefer running one mile over running four, while you only slightly prefer running farther.

The moral of this chapter is that the same sort of situation may arise in

1. See CHRISTENSEN 2007, JOYCE 2007, and WEATHERSON 2007 for further recent references to compromising by splitting the difference.

purely epistemic cases of compromise, cases in which each agent prefers having certain credences over others, where this preference is grounded in purely epistemic concerns. Suppose I strongly prefer assigning .1 credence to a disputed proposition, while you weakly prefer credences around .7. In this kind of case, we may reach a perfect compromise by converging on some shared credence *lower than* .4. Splitting the difference may not constitute a perfect compromise when agents who have different credences also have different epistemic values.

To make this moral precise, we must say how an agent may value certain credences over others, in a purely epistemic sense. I take an agent's *scoring rule* to measure how much she epistemically values various alternative credences she might assign a given proposition. It is natural to suppose that agents should assess just this kind of value, as they judge how much they would prefer certain consensus opinions over others. Using scoring rules, we can develop a natural alternative to the strategy of compromising by splitting the difference: agents may compromise by coordinating on the credences that they collectively most prefer, given their epistemic values.

I have two main aims in this chapter: to develop this alternative strategy, and to argue that this strategy governs how agents should compromise. In §1, I define the notion of a scoring rule and introduce relevant properties of these epistemic value functions. In §2, I develop the alternative strategy for compromising that I defend. In §3, I compare my alternative strategy with the traditional strategy of splitting the difference. I characterize the situations in which the two strategies coincide, and those in which they differ. In §4, I argue that where the strategies do yield different recommendations, compromising by maximizing epistemic value is a reasonable strategy. Finally, in §5, I discuss applications of compromises informed by agents' scoring rules.

## 2.1 Scoring rules

In assigning a particular credence to a proposition, you are estimating its truth value. In a purely epistemic sense, closer estimates are better. Having .8 credence in a truth is better than having .7 credence, and having .9 credence is

better still.[2]

How much better? Different agents may value closer estimates differently. For instance, suppose you care a lot about your credences as they approach certainty. In this case, you may value having .9 credence in a truth much more than having .8 credence, without valuing .6 much more than .5. Or suppose you do not particularly care about having credences that approach certainty, as long as your beliefs are on the right track. In that case, you may equally value .9 and .8 credence in a truth, and equally value .2 and .3 credence in a falsehood.

Facts like these are traditionally modelled by your *scoring rule*, a record of how much you value various estimates of truth values.[3] Formally, a scoring rule $f$ is a pair of functions $f_1$, $f_0$ from $[0,1]$ to $\mathbb{R}$. Intuitively, the first function, $f_1(x)$, measures how much you value having credence $x$ in a proposition that turns out to be true. For instance, if you value having .9 credence in a true proposition much more than having .8 credence, without valuing .6 much more than .5, then the first function $f_1$ of your scoring rule will reflect this preference:

$$[f_1(.9) - f_1(.8)] > [f_1(.6) - f_1(.5)]$$

The second function $f_0(x)$ measures the value of having credence $x$ in a proposition that turns out to be false. For instance, if it makes no difference whether you have .2 or .3 credence in a falsehood, then the second function $f_0$ of your scoring rule will have equal values on these credences:

$$f_0(.2) = f_0(.3)$$

Now we can formally model the statement that closer estimates of truth value are better, i.e. more valuable to you. Closer estimates of the truth value of propositions that turn out to be true are more valuable to you just in case your

2. For further discussion of the notion of credences as estimates, see Jeffrey 1986 and Joyce 1998. For further discussion of the notion of *purely epistemic* value, compare the value of having accurate credences with the more familiar value of having true beliefs, as discussed in Alston 2005 and Lynch forthcoming.

3. Scoring rules were independently developed by Brier 1950 and Good 1952 to measure the accuracy of probabilistic weather forecasts. Savage 1971 uses scoring rules to assess forecasts of random variables, and treats assignments of credence to particular events as a special case of such forecasts. For more recent literature using scoring rules to assess credences, see Gibbard 2006, Joyce 1998, and Percival 2002.

scoring rule $f_1(x)$ is strictly increasing. Closer estimates of the truth value of falsehoods are more valuable just in case $f_0(x)$ is strictly decreasing.

Taken together, your scoring rule and your actual credence in a proposition $p$ determine the *expected epistemic value* of your having a particular credence in $p$. Suppose you actually have credence $\alpha$ in $p$. Then your expected epistemic value of having some other credence $x$ in $p$ is just a weighted sum of the value you assign to having credence $x$ in $p$ if $p$ is true, and to having credence $x$ in $p$ if $p$ is false:

$$\text{EV}(x, \alpha, f_1, f_0) \quad =_{df} \quad \alpha f_1(x) + (1 - \alpha) f_0(x)$$

Your scoring rule says how much value you assign to various credences, including those other than your own. Note that even if you and another agent have the same scoring rule, you may well assign your own credence in $p$ a greater expected epistemic value than you assign her credence, given your own assessment of how likely it is that $p$ is true.

Moreover, you may even assign your own credence in $p$ a greater expected value than you assign *any* alternative credence. If this is always the case, then your scoring rule is *credence-eliciting*. In other words, your scoring rule is credence-eliciting when no matter what credence you assign to a proposition $p$, assigning that credence maximizes the expected value of your credence in $p$. In other words, you always do the best you can by your own lights by assigning $p$ the credence that you do.

Using your scoring rule and your current credence in a proposition, we can calculate the expected value of your having various alternative credences in that proposition. Before applying scoring rules to cases of compromise, I should define one more useful measure: the expected value of your having a particular *credence distribution* over an algebra of many propositions. It is relatively simple to come up with a natural measure of expected value for credence distributions. Roughly, we can compute the expected value of your having several credences in separate propositions by simply summing up the expected value of your having each of those individual credences.

Formally, let $\Gamma$ be your actual credence distribution over some algebra $P$, and let $Q$ be the set of the *atomic propositions* of $P$, i.e. the strongest propositions in $P$ that together form a partition of the space of all possible worlds. Let $D$

be any prospective credence distribution over the $P$ propositions. It is natural to define the expected value of having credence distribution $D$ over $P$ as the sum of the expected values of having the credences assigned by $D$ to the atomic propositions of $P$:

$$\text{EV}(D, \Gamma, f_1, f_0) =_{df} \sum_{q_i \in Q} \text{EV}(D(q_i), \Gamma(q_i), f_1, f_0)$$

In other words, the expected value of a credence distribution is the sum of the expected values of the credences it assigns to atomic propositions.

To take a simple example, suppose you only have opinions about a single proposition $p$. Then your credence distribution is defined over an algebra with atomic propositions $p$ and $\neg p$. The expected value of your having any alternative credence distribution $D$ over this algebra is simply the sum of the expected value of your having credence $D(p)$ in $p$ and credence $D(\neg p)$ in $\neg p$.

## 2.2 Compromise beyond splitting the difference

As long as your scoring rule is credence-eliciting, you prefer your credence in $p$ to any other. That is, as long as your rule is credence-eliciting, your scoring rule will never dictate that you should change your credence in a proposition to another credence with greater expected epistemic value. It is widely accepted that this means that if you are rational, you must have a credence-eliciting scoring rule.[4] Otherwise, your scoring rule could motivate you to raise or lower your credences *ex nihilo*, in the absence of any new evidence whatsoever.

For this reason, how rational agents value alternative credences in a proposition is rarely of practical relevance. Of course, your scoring rule affects how well you think other agents are doing, by affecting the expected value you assign to their credences. But if you are rational, you never value another credence more than your own. So your valuing of non-actual credences will never influence your behavior. For this reason, as long as your scoring rule is credence-eliciting, it will even be hard to tell exactly which scoring rule you have.

However, cases of compromise provide a practical application for the notion of a scoring rule. Scoring rules are practically relevant when agents have to

4. See Oddie 1997, Joyce 1998 and Gibbard 2006 for further discussion.

assess the epistemic value of alternative credences in a proposition. Cases of compromise call for exactly this. Compromising agents must pick a credence to coordinate on. In doing so, they have to assess the epistemic value of alternative credences, if they are to determine which shared credence they most prefer.

This points the way to an alternative strategy for constructing a compromise of agents' opinions: *maximizing expected epistemic value*. Suppose we assign different credences to some proposition $p$, but we must construct a credence that constitutes the compromise of our opinions. The following seems like a natural strategy: choose our consensus credence in $p$ to maximize the average of the expected values that we each assign to alternative credence distributions over the algebra with atomic propositions $p$ and $\neg p$.[5]

For instance, suppose you have credence distribution $A$ and I have credence distribution $B$, and suppose that you score credences with $f_1$ and $f_0$, and I score them with $g_1$ and $g_0$. Then we should compromise by choosing the credence distribution $D$ that maximizes:

$$\textsc{aev}(D, A, f_1, f_0, B, g_1, g_0) \quad =_{df} \quad \frac{1}{2}[\textsc{ev}(D, A, f_1, f_0) + \textsc{ev}(D, B, g_1, g_0)]$$

This is our alternative compromise strategy: rather than just splitting the difference between our credences, we may compromise by maximizing the average of the expected values we give to our consensus credence distribution.[6]

Our alternative compromise strategy yields genuinely alternative recommendations. Disagreeing agents who compromise by maximizing epistemic value rarely end up simply splitting the difference in their credences. Suppose you meet two agents with exactly the same credences, but with different scoring rules. If you were to compromise with each agent individually, you may end up constructing different consensus opinions with each of them, even if your scoring rule is credence-eliciting. This means that if an agent compromises with

5. In order to implement this strategy, we need only assess the epistemic value of credence distributions over simple four-element algebras; the equivalence results I give are restricted to compromises within this relevant class of credence distributions.

6. There are several properties we might wish for in a procedure for aggregating credences; it is notoriously difficult to find a procedure that is satisfactory in all respects. See French 1985 for a discussion of impossibility results, and Genest & Zidek 1986 for a canonical overview and bibliography of the consensus literature. See Shogenji (2007) and Fitelson & Jehle (2007) for discussion of how results about judgment aggregation present challenges for the strategy of splitting the difference.

others by maximizing expected epistemic value, her behavior will carry a lot of information about her scoring rule. In other words, we have discovered that a way in which rational agents value alternative credences can be a matter of practical relevance. Hence cases of compromise give us valuable motivation for studying scoring rules: previously inert differences between scoring rules make for real differences in what compromising agents should do.

## 2.3 Comparing compromise strategies

In some cases, our alternative strategy yields a traditional recommendation. In particular, when agents *share a credence-eliciting scoring rule*, they maximize the average expected value of prospective consensus credence distributions by simply splitting the difference between their credences. For example, suppose we value prospective credences using the *Brier score:*[7]

$$\begin{aligned} f_0(x) &= 1 - x^2 \\ f_1(x) &= 1 - (1-x)^2 \end{aligned}$$

Suppose you have .7 credence in $p$ and I have .1 credence. The Brier score is a credence-eliciting scoring rule. Hence to maximize the average of our expected values for prospective shared credence distributions, we should compromise by giving .4 credence to $p$.

Sometimes agents pursue a perfect compromise, one that is as fair and even as possible. But agents may also pursue an imperfect compromise, one that favors some agents more than others. For example, when an expert and an amateur disagree, they may elect to compromise in a way that favors the expert's original opinion. One traditional strategy for generating an imperfect compromise is to take a weighted arithmetic mean of agents' original opinions.[8] For example, an amateur may give the prior opinion of an expert four times as much weight as his own. Then if the expert initially has .1 credence in $p$ and the amateur has .7 credence, they may compromise at $.8(.1) + .2(.7) = .22$.

7. Strictly speaking, traditional scoring rules measure the *inaccuracy* of a credence, and hence the *disvalue* of having that credence. For instance, Brier 1950 scores credences using the rule $f_0(x) = x^2$, $f_1(x) = (1-x)^2$. For simplicity, I follow Gibbard 2006 in using versions of scoring rules that measure positive epistemic value.

8. For instance, see the discussion of epistemic deference in Joyce 2007.

Our alternative compromise strategy can also be extended to cases of imperfect compromise. In a case of imperfect compromise, agents may maximize a weighted arithmetic mean of the expected values they give to their consensus credence distribution. Furthermore, the above equivalence result extends to cases of imperfect compromise. In the general case: when agents share a credence-eliciting rule, they maximize the weighted average of the expected values they give to their consensus credence distribution when the consensus is the *same weighted average* of their prior credences.[9] It follows that when agents share a credence-eliciting scoring rule, they maximize the exact average of their expected epistemic values by exactly splitting the difference between their prior credences.

Hence in some cases, including cases of imperfect compromise, splitting the difference and maximizing expected value coincide. But it is not hard to see that in principle, these strategies could yield different results. Sometimes agents who disagree about what is *practically* valuable do not maximize their expected utility by splitting the difference between their most preferred outcomes, as in the case where I want to run one mile and you want to run seven. Similarly, agents who disagree about what is epistemically valuable do not always maximize expected epistemic value by splitting the difference in their credences.

For instance, agents with different scoring rules may not maximize expected epistemic value by splitting the difference. Suppose you value prospective credences using the Brier score, and I value them using the following credence-eliciting rule:

$$\begin{aligned} g_0(x) &= x + \log(1-x) \\ g_1(x) &= x \end{aligned}$$

Suppose you have .7 credence in $p$ and I have .1 credence. Even though our scoring rules are each credence-eliciting, our perfect compromise is asymmetric: in constructing a compromise, we maximize our expected epistemic value by choosing a consensus credence of approximately .393, not by splitting the difference.[10]

Even when agents share a scoring rule, they may not maximize expected epistemic value by splitting the difference in their credences. For example,

9. See Corollary of §2.6 for proof.
10. See Example 1 of §2.6 for details.

suppose we both value prospective credences using the following "Brier cubed" score:

$$\begin{aligned} h_0(x) &= 1 - x^3 \\ h_1(x) &= 1 - (1-x)^3 \end{aligned}$$

Suppose you have .7 credence in $p$ and I have .1 credence. Then our perfect compromise is again asymmetric: in constructing a compromise, we maximize our expected epistemic value by choosing a consensus credence of approximately .449.[11]

This value has an interesting property: it is exactly what an agent using the Brier cubed score would assign to $p$ to maximize the expected value of her new credence distribution, if she started with precisely .4 credence in $p$.[12] This is no coincidence: whenever agents share a scoring rule, they maximize the average of their expected values for prospective credence distributions by picking the distribution that has maximal expected value for an agent with their same scoring rule and the average of their credences. Furthermore, this claim extends to a result about weighted averages of agents' expected epistemic values. That is, agents who share a scoring rule can maximize a given weighted arithmetic mean of their expected values for consensus credence distributions by following a straightforward rule; namely, they should choose the credence distribution that a hypothetical agent would most prefer, if she shared their scoring rule, and if her credence in $p$ were just that same weighted arithmetic mean of their actual credences.[13]

This final result incorporates many results given so far. Compromising agents maximize the weighted average of their expected values by choosing the credence distribution preferred by an agent with the same weighted average of their credences. In the special case where compromising agents share a *credence-eliciting* rule, the hypothetical agent with that rule would most prefer her own credence. So agents sharing a credence-eliciting rule should compromise at the weighted average of their actual credence distributions. If the compromising agents pursue a perfect compromise, they maximize their expected epistemic

11. See Example 2 of §2.6 for details.
12. See Example 3 of §2.6 for details.
13. See Theorem of §2.6 for proof.

value by splitting the difference in their credences. But in a number of cases, our alternative compromise strategy comes apart from splitting the difference. It just remains to be argued that in such cases, our alternative compromise strategy is a reasonable one.

## 2.4 Norms governing compromise

In order to understand how compromising agents should be influenced by their epistemic value functions, we must first understand how a single agent should be influenced by her own epistemic values. In the literature on scoring rules, several theorists have addressed the latter question. Our aim is to generalize their suggestions to norms governing multiple agents at once.

It is generally accepted that on pain of irrationality, a single agent must aim to maximize the expected epistemic value of her credences.[14] For example, PERCIVAL 2002 compares epistemic value with practical utility:

> Cognitive decision theory is a cognitive analog of practical decision theory... Bayesian decision theory holds that a rational agent...maximise[s] expected utility. Similarly, Bayesian cognitive decision theory holds that a rational cogniser...maximise[s] expected cognitive utility. (126)

In the same vein, ODDIE 1997 says that scoring rules measure "a (pure) cognitive value which it is the aim of a rational agent, qua pure inquirer, to maximize" (535).

How can we extend this condition to a norm that applies to compromising agents? It is useful to first consider the following single agent case: suppose an evil scientist tells you he is going to perform an operation to change your credence in a certain proposition $p$, which you currently believe to degree .7. The scientist gives you a choice: after the operation, you may have either .6 or .8 credence in $p$. On pain of irrationality, you should choose whichever credence has the greater expected epistemic value. In general, if you are forced to adjust your credence in $p$, so that your new credence satisfies certain constraints, you

14. Or at least she must maximize expected epistemic value, given that her epistemic values are themselves rationally permissible, e.g. credence-eliciting. See PERCIVAL 2002 for further discussion.

should choose the alternative credence with the greatest expected epistemic value for you.

Cases of compromise are relevantly similar. In order to adopt a compromise of their prior opinions, agents are forced to adjust their credences in $p$ so that they satisfy certain constraints. Only in cases of compromise, these constraints are defined extrinsically rather than intrinsically; namely, their adjusted credences must be equal to each other. In this situation, agents should choose the alternative credences with the greatest possible expected epistemic value for them.

In saying precisely how agents should construct a compromise of their opinions, it is useful to see that we face an analogous question when we extend norms governing expected practical utility to cases of practical compromise. Suppose we are deciding where to go for dinner. Based on my credences and practical utilities, I slightly prefer Chinese. Based on your credences and practical utilities, you strongly prefer Indian. Our natural response to the situation is that your stronger preference matters more: we should maximize the average of our expected utilities by choosing Indian. Roughly the same principle governs many reasonable voting systems. Every voter decides which outcome he thinks is most likely to make him the most satisfied. His vote reflects his expected utilities, and the election outcome reflects the average of many agents' expected utilities.

Of course, maximizing average expected utility may not be an ideal decision procedure. But in many situations, maximizing average expected utility is an intuitively reasonable method of deciding what to do when we have different practical utility functions and want to choose a maximally fair action. Epistemic value is the cognitive analog of practical utility. So we need good reason not to use the same method when deciding what to believe, when we have different epistemic utility functions and want to choose a maximally fair credence. In other words, we need good reason to avoid aggregating epistemic preferences in the way we generally aggregate practical preferences. It is not as if we must weigh practical and epistemic utilities in deciding what credence to assign: in epistemology contexts, epistemic values—encoded in scoring rules—are the only values at issue.

One additional reason to prefer the alternative strategy is that agents who

compromise by maximizing the average of their expected epistemic values will never both prefer an alternative shared credence. Other strategies—including most strategies that make no reference to epistemic value—could in principle lead agents to compromise on one credence when there is an alternative that both agents would independently prefer. And that kind of prescription would sit uncomfortably with the norm that every agent should independently aim to maximize her expected epistemic value. If epistemic values should influence an individual in isolation, they should continue to influence her when she compromises with others. The scoring rules strategy for compromising is a natural way of extending accepted norms governing single agents to norms governing compromise.

## 2.5 Applications

In many situations, it is useful to determine not only what we each individually believe, but what we collectively believe. It is reasonable to take what we collectively believe to be what we most value, in a purely epistemic sense. But this identification does not settle how disagreeing peers should update. Suppose that we are disagreeing peers. Let us grant that we should come to have the same beliefs. It is still a further question whether what you should come to believe—and what I should come to believe—is what we collectively believe, in the sense just defined.

If disagreeing peers must adopt a single credence, then adopting what they collectively believe seems like a reasonable choice. But even for those who are skeptical about this potential application of the alternative compromise strategy, cases of compromise are not limited to cases in which disagreeing agents trade in their prior credences for matching ones. Several other epistemic situations call for compromising strategies, including some situations involving many agents, and some involving just one.

Even when disagreeing agents retain their individual credences in a proposition, they may still need to act based on a collective opinion. Disagreeing gamblers may need to decide how much of their shared income to bet on a certain horse. Disagreeing weather forecasters may need to report a single opinion to their employers at the radio station.

In another kind of case, disagreeing agents might not willingly trade in their credences, but might be forced to change them. For instance, suppose an evil scientist tells us he is going to perform an operation to force us to have the same credence in a certain proposition, and that we can choose only what our new credence will be. Intuitively, an individual should ask the scientist for whatever alternative credence has the greatest expected epistemic value. Similarly, we should ask for whatever alternative credence maximizes the average of our expected epistemic values. Recall that if we compromise by splitting the difference in our credences, we might end up compromising at one credence when we would both prefer another. It is hard to see how such a compromise strategy could be rationally obligatory.

Strategies for compromise are also relevant to single agents in complicated epistemic situations. For instance, an agent may use a compromise strategy when updating his credence distribution in light of probabilistic evidence, in the sense of JEFFREY 1968. Intuitively, we can get conflicting probabilistic evidence. Suppose a certain screeching bird looks .8 likely to be a bald eagle, but sounds only .6 likely to be a bald eagle. If you had only seen the bird, it would have been clear how to respond: give .8 credence to the proposition that it is a bald eagle. If you had only heard the bird, it would have been rational to give exactly .6 credence to this proposition. But what credence should you assign in the face of conflicting evidence?

In cases where your visual and aural evidence conflict, your scoring rule may determine how you should combine them. Here is a procedure: suppose there are two agents with exactly your credences, except that one agent updates only on your visual evidence, and the other only on your aural evidence. Determine how those agents would perfectly compromise their opinions. Update by accepting their consensus opinion as your own. To take another example, suppose one weather forecaster says it is .6 likely to rain, and an equally trustworthy forecaster says it is .8 likely. In this case, you may update by maximizing the average of the various expected values you would assign to prospective credence distributions, after updating on information from only one of the forecasters.

Finally, some concerns raised in ELGA 2007 demonstrate that strategies for compromise may also be relevant to agents with imprecise credences. Elga has a negative conclusion in mind: he aims to demonstrate that if an agent

is rational, her credences must be perfectly precise. I am sympathetic with Elga's conclusion, but the arguments needed to establish this claim are more complicated than Elga suggests.

For simplicity, let us suppose that we can represent the belief state of an agent with imprecise credences by a set of probability measures.[15] Elga assumes that an agent with imprecise credences may rationally refuse a bet which is acceptable from the point of view of some but not all probability measures in the set of measures representing her belief state.[16] In other words, if we think of an agent with imprecise credences as if she had a mental committee of agents with precise credences, she may refuse any bet which is unacceptable to any member of her mental committee. But because an agent with imprecise credences may rationally refuse such a large variety of bets, she may rationally refuse a sequence of bets that provides her with an opportunity to win sure money. Foregoing sure money is irrational behavior. So a rational agent must have precise credences.

One could respond to Elga as follows: we should take seriously the suggestion to think of an agent with imprecise credences as if she had a mental committee of agents with precise credences.[17] Such an agent should act in whatever ways a committee should. Elga says that if your mental committee is not unanimously in favor of refusing a bet, "the natural thing to say is that rationality counts the bet as optional for you" (7). But it does not really seem so natural to think that a committee may refuse any bet, as long as just one of its members should prefer to do so. It is more natural to think that in choosing whether to refuse bets, a group should use some compromise strategy to construct a common credence distribution, and then act on the basis of this constructed opinion. Hence we should not accept Elga's premise that a rational agent with imprecise credences may rationally refuse any bet which is acceptable from the point of view of only some members of her mental committee. By acting on a compromise of her mental committee's opinions, a rational agent could be

15. See for instance TINTNER 1941, SMITH 1961, LEVI 1980, JEFFREY 1983, JOYCE 2005, VAN FRAASSEN 2006.

16. Several advocates of imprecise credences endorse this "conservative" betting strategy. See WILLIAMS 1976, KAPLAN 1996, WALLEY 1991, and an extensive catalog of ongoing research at `http://www.sipta.org`.

17. One might also respond that an agent's refusing certain bets constrains what other bets she may rationally refuse. ELGA 2007 addresses this response. I will set it aside here.

rationally compelled to accept sequences of bets that let her win money, come what may.

Elga briefly considers the possibility that agents with imprecise credences may still be rationally obliged to accept or refuse any particular bet. He suggests that on such a proposal, "the interval-valued probabilities do little, since they "collapse down" to ordinary point-valued probabilities when it comes to imposing constraints on rational action" (8). But this response is too quick. For instance, suppose that an agent with imprecise credences updates by "pointwise" conditionalization. In other words, suppose that on receiving some evidence, she updates the set of probability measures representing her belief state by conditionalizing each measure on the evidence received. Furthermore, suppose she has a single credence-eliciting scoring rule, and so acts according to a compromise of the opinions of her mental committee members, all of whom share a credence-eliciting scoring rule. Then she will act according to an arithmetic mean of her committee members' credence distributions.[18] But taking an arithmetic mean of several distributions does not commute with conditionalizing those distributions on a given proposition. So this kind of agent with imprecise credences will act at each stage as if she has a precise credence distribution, but will not act over time as if she has a precise credence distribution that she updates by conditionalization.

This result contradicts Elga's suggestion that if an agent with imprecise credences faces stringent obligations to accept or reject particular bets, her credences may as well be precise "when it comes to imposing constraints on rational action." But the result is friendly towards Elga's overall conclusion. I have argued that as long as an agent has a single scoring rule and updates by "pointwise" conditionalization, she will not act as if she is updating a single precise credence distribution by conditionalization. For instance, she will be subject to diachronic Dutch Books.[19] If that means she is irrational, then we have a limited

18. Here I take an agent's belief state to be represented by a finite set of measures. In this case, the above result follows from the Corollary proved in §2.6. Representing an agent's belief state by an infinite set of measures creates an additional problem: how to parameterize the space of her committee's distributions when calculating their arithmetic mean. This problem resembles Bertrand's paradox. Some theorists treat credences as imprecise chiefly in order to *avoid* paradoxes of this kind. Such theorists have an additional reason to reject the betting strategy currently under consideration.

19. See TELLER 1973, LEWIS 1999.

version of the conclusion Elga is after, which is a step towards demonstrating the irrationality of imprecise credences.

## 2.6 Proofs

**Example 1.** Running the following *Mathematica* notebook verifies that agents with different credence-eliciting scoring rules may not maximize the average of their expected epistemic values in an alternative credence distribution by splitting the difference in their credences.

```
b1[x_] := 1 - (1 - x)^2; b0[x_] := 1 - x^2
l1[x_] := x; l0[x_] := x + Log[1 - x]
ev[x_, m_, f1_, f0_] := (m*f1[x]) + (1 - m)*f0[x]
CDev[x_, m_, f1_, f0_]
  := ev[x, m, f1, f0] + ev[(1 - x), (1 - m), f1, f0]
avgCDev[x_, m_, m1_, m0_, n_, n1_,  n0_]
  := .5 (CDev[x, m, m1, m0] + CDev[x, n, n1, n0])
Maximize[{avgCDev[x, .7, b1, b0, .1, l1, l0], 0 <= x <= 1}, x]
```

For an agent with the Brier score and credence .7 in $p$, and an agent with scoring rule $l_1$, $l_0$ and credence .1 in $p$, having approximately credence .392965 in $p$ maximizes expected epistemic value:

```
{0.924402, {x -> 0.392965}}.
```

**Example 2.** Running the previous notebook with the following additions verifies that agents with non-credence-eliciting scoring rules may not maximize the average of their expected epistemic values in an alternative credence distribution by splitting the difference in their credences.

```
c1[x_] := 1 - (1 - x)^3; c0[x_] := 1 - x^3
Maximize[{avgCDev[x, .7, c1, c0, .1, c1, c0], 0 <= x <= 1}, x]
```

For agents who share the Brier cubed score and have credences .7 and .1 in

$p$, having approximately credence .44949 in $p$ maximizes expected epistemic value:

```
{1.75755, {x -> 0.44949}}.
```

**Example 3.** Running the previous notebook with the following addition verifies that the credence distribution preferred by agents sharing a scoring rule in Example 2 is the credence distribution a hypothetical agent would prefer, if she shared that same scoring rule and had the arithmetic mean of their credences:

```
Maximize[{CDev[x, .4, c1, c0], 0 <= x <= 1}, x]
```

For an agent with the Brier cubed score and credence .4 in $p$, having approximately credence .44949 in $p$ maximizes expected epistemic value:

```
{1.75755, {x -> 0.44949}}.
```

**Theorem.** If a finite number of agents each use a scoring rule that differs by no more than a constant from a single "shared" scoring rule, then they maximize the weighted average of the expected values they give to a consensus credence distribution by choosing the distribution that a hypothetical agent with their shared scoring rule would prefer, if she were to have that same weighted average of their credences.[20]

**Proof.** Let us say there are $n$ agents, and that for all $i \in [1, n]$, the $i$th agent uses scoring rule $g_1^i$, $g_0^i$ and has credence distribution $\delta_i$ over the algebra with atomic propositions $p$ and $\neg p$.

There exist coefficients $c_i$ such that $1 = \sum_{i=1}^{n} c_i$ and such that the following is the weighted average of the expected epistemic values that the compromising agents give to a consensus credence distribution $D$ over the algebra with atomic

20. This result is restricted to credence distributions over four-element algebras, since determining how agents should compromise on such simple credence distributions is sufficient to determine how agents should compromise when they assign different credences to a single proposition. See footnote 5.

propositions $p$ and $\neg p$:

$$\begin{aligned}
&\textsc{waev}(D, \delta_1, g_1^1, g_0^1, \ \dots\ , \delta_n, g_1^n, g_0^n) \\
&= \sum_{i=1}^{n} c_i \textsc{ev}(D, \delta_i, g_1^i, g_0^i) \\
&= \sum_{i=1}^{n} [c_i \textsc{ev}(D(p), \delta_i(p), g_1^i, g_0^i) + c_i \textsc{ev}(D(\neg p), \delta_i(\neg p), g_1^i, g_0^i)] \\
&= \sum_{i=1}^{n} c_i \textsc{ev}(D(p), \delta_i(p), g_1^i, g_0^i) + \sum_{i=1}^{n} c_i \textsc{ev}(D(\neg p), \delta_i(\neg p), g_1^i, g_0^i)
\end{aligned}$$

By supposition, there is a scoring rule $f_1$, $f_0$ such that for any $i \in [1, n]$, there are constants $k_i, l_i \in \mathbb{R}$ such that $g_1^i = f_1 + k_i$ and $g_0^i = f_0 + l_i$. So we can reduce the first summand as follows:

$$\begin{aligned}
&\sum_{i=1}^{n} c_i \textsc{ev}(D(p), \delta_i(p), g_1^i, g_0^i) \\
&= \sum_{i=1}^{n} c_i [\delta_i(p)(f_1(D(p)) + k_i) + (1 - \delta_i(p))(f_0(D(p)) + l_i)] \\
&= \sum_{i=1}^{n} c_i \delta_i(p) f_1(D(p)) + \sum_{i=1}^{n} c_i (1 - \delta_i(p)) f_0(D(p)) + \sum_{i=1}^{n} c_i [\delta_i(p) k_i + (1 - \delta_i(p)) l_i] \\
&= \sum_{i=1}^{n} c_i \delta_i(p) f_1(D(p)) + (1 - \sum_{i=1}^{n} c_i \delta_i(p)) f_0(D(p)) + \sum_{i=1}^{n} c_i [\delta_i(p) k_i + (1 - \delta_i(p)) l_i]
\end{aligned}$$

This function is simply the sum of the constant $\sum_{i=1}^{n} c_i [\delta_i(p) k_i + (1 - \delta_i(p)) l_i]$ and the expected value that an agent with the scoring rule $f_1$, $f_0$ gives to credence $D(p)$ in $p$, when she has credence $\sum_{i=1}^{n} c_i \delta_i(p)$ in $p$.

Similarly, the second summand, $\sum_{i=1}^{n} c_i \textsc{ev}(D(\neg p), \delta_i(\neg p), g_1^i, g_0^i)$, is the sum of the constant $\sum_{i=1}^{n} c_i [\delta_i(\neg p) k_i + (1 - \delta_i(\neg p)) l_i]$ and the expected value that an agent with the scoring rule $f_1$, $f_0$ gives to having credence $D(\neg p)$ in $\neg p$, when she has credence $\sum_{i=1}^{n} c_i \delta_i(\neg p)$ in $\neg p$.

The hypothetical agent has credence $\sum_{i=1}^{n} c_i\delta_i(\neg p)$ in $\neg p$ just in case her credence $C(p)$ in $p$ is as follows:

$$\begin{aligned} C(p) &= 1 - C(\neg p) \\ &= 1 - \sum_{i=1}^{n} c_i\delta_i(\neg p) \\ &= \sum_{i=1}^{n} c_i - \sum_{i=1}^{n} c_i(1 - \delta_i(p)) \\ &= \sum_{i=1}^{n} c_i\delta_i(p). \end{aligned}$$

So the second summand is the sum of a constant term and the expected value that an agent with the scoring rule $f_1$, $f_0$ gives to having credence $D(\neg p)$ in $\neg p$, when she has credence $\sum_{i=1}^{n} c_i\delta_i(p)$ in $p$.

Hence the initial value $\textsc{waev}(D, \delta_1, g_1^1, g_0^1, \ldots, \delta_n, g_1^n, g_0^n)$ is the sum of a constant term and two expected values: the expected value that an agent with the scoring rule $f_1$, $f_0$ gives to credence $D(p)$ in $p$ and the expected value that she gives to credence $D(\neg p)$ in $\neg p$, when she has credence $\sum_{i=1}^{n} c_i\delta_i(p)$ in $p$. In other words, $\textsc{waev}(D, \delta_1, g_1^1, g_0^1, \ldots, \delta_n, g_1^n, g_0^n)$ is the sum of a constant term, and the expected value that an agent with the scoring rule $f_1$, $f_0$ gives to the *credence distribution* $D$, when she has credence $\sum_{i=1}^{n} c_i\delta_i(p)$ in $p$.

Since $\textsc{waev}(D, \delta_1, g_1^1, g_0^1, \ldots, \delta_n, g_1^n, g_0^n)$ and the expected value of an agent with credence $\sum_{i=1}^{n} c_i\delta_i(p)$ in $p$ differ only by the addition of a constant term, these functions are maximized at the same values. So agents maximize the weighted average of the expected values they give to a consensus credence distribution, i.e. $\textsc{waev}(D, \delta_1, g_1^1, g_0^1, \ldots, \delta_n, g_1^n, g_0^n)$, by choosing the distribution that a hypothetical agent with their shared scoring rule would prefer, if she were to have that same weighted average of the compromising agents' credences, i.e. $\sum_{i=1}^{n} c_i\delta_i(p)$. □

**Corollary.** If a finite number of agents share a credence-eliciting scoring rule, then they maximize the weighted average of the expected values they give to their consensus credence distribution by choosing the credence distribution that assigns $p$ that same weighted average of their credences.

**Proof.** If agents share a scoring rule, then they each use a scoring rule that differs by no more than a constant (namely, 0) from a single "shared" scoring rule. So by the above Theorem, the agents maximize the weighted average of the expected values they give to a consensus credence distribution by choosing the distribution that a hypothetical agent with their shared scoring rule would prefer, if she were to have that same weighted average of the compromising agents' credences.

If the shared scoring rule is credence-eliciting, then the hypothetical agent will prefer her own credences in $p$ and $\neg p$ over any other credences in those propositions. So she will prefer her own credence distribution over any other. Hence agents with a shared credence-eliciting rule maximize the weighted average of the expected values they give to their consensus credence distribution by choosing the credence distribution that assigns $p$ that same weighted average of their credences. □

## 3 Updating as communication

On many traditional theories of belief, your belief state is represented by an assignment of credences to propositions, or sets of possible worlds. If you are rational, your credence distribution will be a probability measure. Traditional theories of belief fit with a standard Bayesian theory of rational belief change: on learning a proposition, you must update your belief state by conditionalizing your credence distribution on the proposition you learn. That is, you must update by assigning 0 credence to those worlds incompatible with what you learn, and re-normalizing your credence distribution over the remaining worlds.

Following Quine 1969, Lewis 1979 argues that we should instead represent your belief state by an assignment of credences to sets of *centered worlds*: world-time-individual triples. For instance, if you have .5 credence that it is 3:00pm, your belief state should be represented by a measure that assigns .5 to the set of centered worlds with that time coordinate. Unlike traditional theories of belief, Lewis's theory does not fit with a standard Bayesian theory of rational belief change. For instance, Bayesian conditionalization preserves certainties. If you update by conditionalizing on the set of centered worlds you learn, it follows that if you are ever certain that it is 3:00pm, you must always remain certain that it is 3:00pm. But clearly this is not what rationality requires. If we agree with Lewis about how to represent belief states, we must develop another set of principles governing rational belief change.

In this chapter, I develop a procedure for rationally updating credence distributions over sets of centered worlds. I argue that rational updating can be factored into two steps. Roughly speaking, in forming an updated credence distribution, you must first use information you recall from your previous self to form a hypothetical credence distribution, and then change this hypothetical distribution to reflect information you have genuinely learned as time has passed. In making this proposal precise, I argue that your recalling information from your previous self resembles a familiar process: agents' gaining information from each other through ordinary communication.

The updating procedure I develop relies on relationships between two kinds of sets of centered worlds: *de se* and *de dicto* propositions. I will define *de dicto* propositions to be boring sets of centered worlds: sets of world-time-individual triples such that if one triple is in the set, so is every other triple which shares its world coordinate. *De se* propositions are sets of centered worlds that are not *de dicto* propositions. *De dicto* propositions are entirely about what the world is like, while *de se* propositions are also about where you are in the world. In §1, I make some observations about how *de se* contents of attitudes are related to *de dicto* propositions. In §2, I use these observations to solve a puzzle about imagination.

The discussion in §1-2 provides the foundation for a unified theory of communicating and updating beliefs. In §3, I describe how agents communicate *de se* beliefs. In §4, I argue that rational updating begins with a similar process. In §5, I introduce the rest of a complete procedure for rationally updating credences in *de se* propositions. Finally, in §6, I apply my theory to a particular case. The case fits well with my theory, but presents a problem for the theory of updating given in TITELBAUM 2008. This problem for Titelbaum exposes a difference between his theory and mine, and highlights an important feature of rational updating that any successful updating procedure must recognize.

## 3.1 *De se* and *de dicto* contents

In giving a theory of how you should update your *de se* beliefs, it is helpful to understand how the contents of those beliefs are related to various *de dicto* propositions. I will focus on the following observation: given a *de se* proposition, there is a *de dicto* proposition that is equivalent with that *de se* proposition, given what you believe. In more precise terms: given a *de se* proposition, there is a *de dicto* proposition such that for any centered world compatible with what you believe, that centered world is in the former proposition just in case it is in the latter.

Semantic theories of attitude ascriptions can help us find *de dicto* propositions equivalent with contents of *de se* attitudes. One helpful claim generally accepted by semanticists is that speakers can use first-person indexicals to self-

ascribe attitudes with *de dicto* contents.[1] For example, suppose Kaplan sees himself in a mirror, without realizing that he is seeing himself. Looking at the mirror, Kaplan sees that his pants are on fire, without realizing that his own pants are on fire. In recounting his experience, suppose Kaplan says:

(1) I expected that I would be rescued.

Kaplan can truly utter (1), even though he was not aware of being in danger when he looked at the mirror. In this respect, (1) differs from (2):

(2) I expected to be rescued.

Unlike (1), reports such as (2) can be true only if the ascribee has a self-directed attitude. There is general consensus about the best way to model this difference: we say reports such as (2) can only ascribe attitudes with *de se* contents, while reports such as (1) can ascribe attitudes with *de dicto* contents.[2] The content of the expectation ascribed by (2) must be a set of centered worlds where the center is rescued. But the content of the expectation ascribed by (1) can be a set of centered worlds that is characterized not by any property of the center, but by some property of the person Kaplan sees. So (1) and (2) can ascribe expectations with different contents, and that is why these ascriptions can have different truth conditions.

Our semantic theory says that (1) can ascribe a *de dicto* attitude. Kaplan believes that the person he sees is not himself, so the content of the *de dicto* attitude that (1) ascribes is not equivalent with the content of the *de se* attitude that (2) ascribes, given what he believes. But normally when a speaker utters (1) and (2), the content of the *de dicto* attitude that (1) ascribes is equivalent with the content of the *de se* attitude that (2) ascribes, given what she believes. These contents are still distinct propositions. In particular, only one is a *de se* proposition. But the centered worlds at which they differ in truth value are not among the worlds compatible with what the speaker believes.

1. Morgan 1970 and Lakoff 1972 were among the first to highlight third-personal readings of embedded first-person pronouns. In setting up my examples, I use a case developed by Kaplan in the late 1970's and familiar from Kaplan 1989.
2. For more detailed semantic proposals starting from this point of consensus, see Chierchia 1989, von Stechow 2002, von Fintel 2005, Anand 2006, and Ninan 2008. For a helpful overview of many such proposals, see Ninan 2009.

The moral here is that you can always have *de dicto* beliefs about yourself, just as you can have *de dicto* beliefs about any other person. In cases without identity confusion, you have a third-personal way of thinking about yourself. This way of thinking about yourself is what gives rise to the *de dicto* attitudes that you use first-person indexicals to self-ascribe. The contents of these attitudes are equivalent with the contents of your *de se* attitudes, given what you believe. Furthermore, even Kaplan has a normal third-personal way of thinking about himself, in addition to having thoughts caused by his image in the mirror. The contents of his resulting *de dicto* attitudes are equivalent with the contents of his *de se* attitudes, given what he believes. So whether or not you are in an identity confusion case, there is a *de dicto* proposition that is equivalent with any given *de se* proposition, given what you believe.

Once we recognize that we use first-person indexicals to ascribe attitudes with *de dicto* contents, we can see that an even stronger moral holds: there is a *de dicto* proposition that is equivalent with any given *de se* proposition, given merely what you believe with certainty. Even if Kaplan started to believe that he himself was the guy whose pants were on fire, he could always have some shred of doubt about this conclusion. Contrast this with your immediate conviction, on uttering (1) and (2) in a normal case, that if one expectation is satisfied then the other will be. Similarly, on uttering (3), you cannot doubt that your expectation is about yourself:

(3) I expect that I will be rescued.

In just this sense, you are always certain about which person is yourself. Given what you believe with certainty, the contents of your *de dicto* beliefs about that person will be equivalent with the contents of your *de se* beliefs about yourself. Or in fewer words: your *de dicto* beliefs about that person will be equivalent with your *de se* beliefs about yourself.

First-person indexicals are not the only way to ascribe *de dicto* attitudes equivalent with your *de se* attitudes. Suppose Kaplan introduces a name for himself:

(4) Let 'Dr. Demonstrative' name myself.

On uttering (4), Kaplan can be certain that he himself is Dr. Demonstrative. So

given what Kaplan believes with certainty, (5) and (6) ascribe expectations with equivalent contents:

(5) I expect to be rescued.

(6) I expect that Dr. Demonstrative will be rescued.

On any semantic theory of ascriptions, Kaplan uses (6) to ascribe an expectation with a *de dicto* content. In other words, Kaplan uses 'Dr. Demonstrative' to think about himself in a normal third-personal way. So he may use this name to self-ascribe *de dicto* attitudes equivalent with his *de se* attitudes, given what he believes with certainty.

To sum up so far: not all of your *de dicto* attitudes about yourself are equivalent with your *de se* attitudes, given what you believe with certainty. For instance, Kaplan could always have some shred of doubt about whether he is the man whose pants are on fire, or even about whether he is David Kaplan. So the *de se* expectation that he ascribes using (5) is not equivalent with the expectations he ascribes using (7) and (8):

(7) I expect that the man whose pants are on fire will be rescued.

(8) I expect that David Kaplan will be rescued.

But other *de dicto* attitudes about yourself are equivalent with your *de se* attitudes, given what you believe with certainty. For instance, the expectation ascribed in (5) is normally equivalent with expectations ascribed using first-person indexicals (as in (3)), and always equivalent with expectations ascribed using names introduced with reflexive expressions (as in (6)).

The same results hold for your *de se* attitudes about your temporal location. Suppose that it is 3:00 and that Kaplan is perfectly aware of the time. Consider the following ascriptions:

(9) Kaplan believes that it is 3:00.

(10) Kaplan believes that now is 3:00.[3]

The beliefs ascribed in (9) and (10) may appear to have the same content. But we

3. For continuity with earlier examples, I use 'now' as a referring expression. Those who object may replace 'now is 3:00' with 'it is now 3:00' without consequence.

can distinguish each content by asking whether Kaplan will continue to have a belief with that content as time passes:

(11) #At 5:00, Kaplan will still believe that it is 3:00.

(12) At 5:00, Kaplan will still believe that now is 3:00.

Our intuitive judgment is that as time passes, Kaplan loses the belief ascribed in (9) and retains the belief ascribed in (10). This is reason to think that (9) and (10) ascribe beliefs with different contents. Intuitively, Kaplan should give up the belief ascribed in (9) when his temporal location changes, while the belief ascribed in (10) is about a particular fixed time whose characteristics do not depend on his temporal location. In other words, (9) ascribes a belief with a *de se* content while (10) ascribes a belief with a *de dicto* content. But these contents are equivalent, given what Kaplan believes with certainty at 3:00. In addition to thinking about yourself from an impersonal perspective, you can think about your temporal location from an atemporal perspective. Just as with your impersonal thoughts about yourself, the resulting *de dicto* attitudes are equivalent with your *de se* attitudes, given what you believe with certainty.

### 3.2 Two ways of imagining

Finding *de dicto* propositions equivalent with *de se* attitude contents is not only useful for theorizing about communicating and updating. It can help us solve a puzzle about imagination and other similar attitudes. This puzzle provides independent motivation for the ideas I have introduced.

Suppose that it is 3:00 and you are teaching class, and while you are teaching, I ask you to imagine that it is 5:00. There are two very different ways you might respond. For instance, you might play along by saying either of the following:

(13) Then I am in my kitchen, starting to make dinner.

(14) Then my watch is wrong, and all of us must be strangely confused to be here so much later than usual.

Once you decide to respond in one of these ways, it is clear how you should

go on with what you are imagining. Either you imagine that two hours have passed and your day has proceeded normally, or you imagine that someone has played a practical joke on you and your students. These responses involve very different kinds of imaginary scenarios. The acceptability of either response raises a puzzle: what distinguishes these two ways of imagining that it is 5:00?

In both cases, when I ask you to imagine that it is 5:00, you comply by imagining a certain *de se* proposition. In particular, all centered worlds compatible with what you imagine are in the set of centered worlds whose time coordinate is 5:00. But what you imagine in each case is distinguished by whether you also imagine a certain *de dicto* proposition.

Outside the pretense, you actually believe the *de se* proposition that it is 3:00. I can ascribe this *de se* belief to you by saying:

(15) You believe that it is 3:00.

Furthermore, you have a *de dicto* belief equivalent with this *de se* belief, given the propositions that you actually believe with certainty. Namely:

(16) You believe that now is 3:00.

The content of the belief ascribed in (16) is central to our solution of the puzzle. The different ways of imagining that it is 5:00 are fundamentally separated by whether what you imagine is consistent with what you actually believe. In the case where you imagine as in (13), you not only imagine the *de se* proposition that it is 5:00, but also the *de dicto* content of the belief ascribed in (16). In the case of (14), this *de dicto* proposition is not part of what you imagine. In other words, there is an extra constraint on the worlds compatible with what you imagine in (13): the *de dicto* content of the belief ascribed in (16) holds in all these worlds.

Our natural responses to (13) and (14) support my characterization of the difference between these ways of imagining. For instance, it is natural to say that when you accept (13), you are imagining that *some time has passed.* If you are imagining that the actual current time has already passed, you may freely imagine that it is 5:00, while imagining that you correctly identified the actual current time as 3:00. In this case, your *de dicto* belief that the actual current time is 3:00 is true at worlds compatible with what you imagine. By contrast,

it is natural to say that when you accept (14), you are imagining that *the actual current time is not what you thought it was*. In this case, your *de dicto* belief that the actual current time is 3:00 is not true at worlds compatible with what you imagine.

The same puzzle arises for several attitudes besides imagining. For example, there are two natural ways to suppose the *de se* proposition that it is 5:00, corresponding to two indicative conditionals:

(17) If it is 5:00, then I am in my kitchen, starting to make dinner.

(18) If it is 5:00, then my watch is wrong, and all of us must be strangely confused to be here so much later than usual.

Here the puzzle is to say why both of these very different conditionals can be acceptable.

Let us agree with RAMSEY 1931 that 'if $p$, would $q$' is acceptable to those who accept $q$ after "adding $p$ hypothetically to their stock of knowledge" (248). Both (17) and (18) can be acceptable because there are different ways to add the *de se* proposition that it is 5:00 to your stock of knowledge. In particular, as you suppose that it is 5:00, you may or may not continue to accept the *de dicto* content of the belief ascribed in (16). If you retain your *de dicto* belief that the actual current time is 3:00, you will accept the consequent of (17). If you give up your *de dicto* belief, you will accept the consequent of (18).

So far I have distinguished ways of imagining and supposing centered contents. The distinctions I have drawn are related to the distinction between belief updating and belief revision often cited in literature on *de se* belief change.[4] In order to accept the consequent of (17), you must update on the antecedent as if some time had passed. In order to accept the consequent of (18), you must instead revise your current beliefs. In both updating and revising, you give up some *de se* beliefs. Updating and revising are distinguished by whether you also give up certain *de dicto* beliefs that your old *de se* beliefs were equivalent with. If you retain your *de dicto* beliefs, you are updating. If you give them up, you are revising.

Retaining *de dicto* beliefs equivalent with your *de se* beliefs is what unifies

4. For instance, see KATSUNO & MENDELZON 1991, 'On the Difference Between Updating a Knowledge Base and Revising it'.

several attitudes: imagining as in (13), supposing as in (17), and updating rather than revising. I hope to have forestalled the objection that ascriptions like (16) ascribe only trivial *de dicto* beliefs, by arguing that whether you retain such *de dicto* beliefs grounds substantive differences in ways of imagining and supposing propositions. I also hope to have forestalled the objection that ascriptions like (16) in fact ascribe *de se* beliefs, since you imagine the same *de se* contents in (13) and (14), while you imagine the content of the belief ascribed in (16) only in the former case.

To sum up, our puzzle about imagining gives us reason to think that there are non-trivial *de dicto* beliefs equivalent with your *de se* beliefs, given what you believe with certainty. Specifically, positing such *de dicto* beliefs yields a simple and intuitive solution to our puzzle. In what follows, I give another reason to accept *de dicto* beliefs self-ascribed by indexicals: as I will argue, they play an important role in a simple unified theory of *de se* communication and updating.

## 3.3 Learning from other agents

In §4-5, I develop a theory of how agents should maintain and modify their *de se* beliefs as time passes. On this theory, part of updating resembles another instance of the transmission of centered information: interpersonal communication. Communicating agents may exchange beliefs, even though they distribute their credence over entirely disjoint centered propositions, namely sets of worlds with distinct person coordinates. Similarly, an agent may retain beliefs over time, even though at different times, she distributes her credence over sets of worlds with distinct time coordinates.

Lewis says that many belief contents are *de se* propositions. But these *de se* propositions cannot always be what is conveyed in communication. For example, suppose Kaplan believes that his own pants are on fire, and when he tells his sister what he believes, she comes to believe just this same centered proposition. Then his sister would come to believe that her own pants were on fire. This is the same centered proposition that Kaplan believes. But obviously, it is not the information that Kaplan should have conveyed to his sister, in telling her what he believed. Instead she should have come to believe some other *de se* propositions, such as the set of centered worlds where the center has a brother

whose pants are on fire.

The same goes for the transmission of centered information across times. Suppose I express one of my beliefs by saying 'it is Monday' and one day later I remember this belief. Then I should not come to self-ascribe the property of being located on Monday, but the property of being located on Tuesday.

These examples illustrate a *prima facie* tension between two intuitive ideas. On the one hand, we may favor a Stalnakerian "package delivery" model of communication, on which what I believe is what you come to believe when I communicate my beliefs. On the other hand, Lewis suggests that I believe *de se* propositions. But when I communicate my beliefs, you do not come to believe the same *de se* propositions that I believe. Instead you come to believe other *de se* propositions, ones that I don't believe.

It is not hard to resolve this tension with notions we already have at hand. There is something that Kaplan believes, that he tells his sister, and that his sister comes to believe. It is the *de dicto* content of the belief Kaplan could self-ascribe by saying:

(19) I believe that my pants are on fire.

In other words, Kaplan uses 'my pants are on fire' to convey a *de dicto* content equivalent with the *de se* proposition that his own pants are on fire, given what he believes with certainty. In coming to believe this proposition, Kaplan's sister does not come to believe that her own pants are on fire. Of course, she may acquire several *de se* beliefs of her own. For instance, she may infer that she herself has a brother whose pants are on fire. But the "delivered package" of the Stalnakerian model is a *de dicto* proposition. Just as we can use indexicals to self-ascribe *de dicto* beliefs, we can use indexicals to convey *de dicto* information.

This theory fits with the Lewisian framework, while respecting our intuitions about the identity conditions of contents conveyed in conversation. STALNAKER 2008 worries that the Lewisian framework conflicts with our intuitions about individuating contents:

> Lewis's account distinguishes contents that ought to be identified. If Rudolf Lingens tells you that he is sad, or that he is Rudolf Lingens, and you understand and accept what he says, then it seems that the information you acquire is the same information he imparted. (1)

But Lewis can accommodate this intuition, while still taking belief contents to be sets of centered worlds. If Lingens tells you that he is sad, he conveys a *de dicto* proposition equivalent with the content of his *de se* belief that he himself is sad, given what he believes with certainty. This proposition is something that Lingens believes, that he conveys, and that you come to believe. Our judgment that we should identify what you and Lingens believe reflects the fact that you both believe this *de dicto* proposition. Stalnaker also worries:

> [Lewis] identifies contents that ought to be distinguished. What I believe when I believe that I was born in New Jersey is something about myself, something different from what my fellow New Jersey natives believe about themselves. What I tell the waiter when I tell him that I will have the mushroom souffle is different from what you tell the waiter if you decide to have the same thing. (1)

But Lewis may respond that when Stalnaker believes that he was born in New Jersey, he believes a *de dicto* proposition equivalent with the content of his *de se* belief that he himself was born there, given what he believes with certainty. His fellow New Jersey native believes a different *de dicto* proposition. Our judgment that we should distinguish what Stalnaker and his fellow New Jersey native believe reflects the fact that they believe different *de dicto* propositions. Similarly, our judgment that we should distinguish what you and Stalnaker tell the waiter reflects the fact that you convey different *de dicto* propositions to the waiter, even if you use the same indexicals when you order.

This discussion suggests a simple theory of the role your *de se* beliefs play in communication. Each *de se* proposition you believe is equivalent with some *de dicto* proposition, given what you believe with certainty. This kind of *de dicto* proposition is something you convey to your audience, and something they come to believe. Furthermore, your audience already has some *de se* beliefs about their relation to you. So they also come to believe some *de se* propositions: the consequences of their standing *de se* beliefs and their acquired *de dicto* information.[5]

Suppose we are standing in a line. I see that I am just behind you, but I have no idea how many people are ahead of you. Suppose you believe a *de se*

5. This is a theory of how agents normally communicate. See EGAN 2005 for arguments that speakers use epistemic modals to directly convey *de se* propositions.

proposition: that you yourself are fourth in line. This proposition is equivalent with some *de dicto* proposition, given what you believe with certainty. If you say 'I am fourth in line' to me, then this kind of *de dicto* proposition is something that you convey to me, and something that I come to believe. Furthermore, I already have some *de se* beliefs about my relation to you: that I myself am just behind you in line. So I also come to believe a *de se* proposition: that I myself am fifth in line. So when we communicate, I gain *de se* beliefs: not your beliefs, but the consequences of my standing *de se* beliefs and my acquired *de dicto* information.

In §1, I argued that we embed (20) in ascriptions of *de se* attitudes and (21) in ascriptions of *de dicto* attitudes:

(20) It is 3:00.

(21) Now is 3:00.

The theory of communication I have outlined suggests that we nevertheless use (20) and (21) to convey the same *de dicto* information.[6] So while *de se* propositions may serve as compositional semantic values and the contents of our attitudes, *de dicto* propositions are the common currency in which we convey information to each other.

These claims about (20) and (21) invoke a more general thesis about language: the Dummettian distinction between assertoric content and ingredient sense.[7] Dummett says that the assertoric content of a sentence may not determine its ingredient sense. Or in more familiar terms, what is said by a sentence may not determine what it contributes to the truth conditions of sentences in which it is embedded. I have set aside some important questions, such as whether the content conveyed by an utterance is "what is said" by that utterance, and whether it is a semantic or pragmatic fact that utterances of (20) and (21) convey the same *de dicto* information. But my discussion suggests a moral in the spirit of the Dummettian distinction: the *de dicto* proposition conveyed by a sentence like (20) or (21) may not determine whether that sentence contributes

6. In particular, I take it that you convey the *de dicto* content of a belief that you would use an indexical to self-ascribe, such as the *de dicto* beliefs ascribed by (3) and (10).

7. See DUMMETT 1991 for exposition of the distinction and STANLEY 1997 for an interpretation relevant to what I say here.

a *de dicto* or *de se* proposition to the truth conditions of an attitude ascription.

## 3.4 Learning from your previous self

Giving a theory of how agents with *de se* beliefs communicate illuminates how agents maintain and modify their *de se* beliefs over time. The model of updating I will give relies on an intuitive notion of *genuine learning*. Everyone recognizes that as you sense that time is passing, you should change your credences to reflect your awareness of your changing temporal location. And your opinions about exactly how much time has passed should influence how you update. But ordinarily as time passes, you are not merely sitting in a black box, keeping track of the minutes as they pass by. You have experiences that make you more informed than your previous self, imposing novel constraints on your credences. In other words, you genuinely learn information. In what follows, I will take for granted the distinction between updating in a black box, and updating as you genuinely learn information.

In black box updating, you form beliefs on the basis of information you get from your previous self. Getting information from your previous self is just like getting information from other agents. Each *de se* proposition you used to believe is equivalent with some *de dicto* proposition, given what you used to believe with certainty. This kind of *de dicto* proposition is something you can currently believe. Furthermore, you currently have some *de se* beliefs about your relation to your previous self. So you can also currently believe some *de se* propositions: the consequences of your current *de se* beliefs and your old *de dicto* information.

Suppose you used to believe a *de se* proposition: that it was the fourth of the month. This proposition is equivalent with some *de dicto* proposition, given what you used to believe with certainty. This kind of *de dicto* proposition is something you can currently believe. Furthermore, you currently have some *de se* beliefs about your relation to your previous self: that your current self is located one day later. So you can also currently believe a *de se* proposition: that it is the fifth of the month. Just as an agent may have certain *de se* beliefs once she acquires *de dicto* beliefs from other agents, you may have certain *de se* beliefs once you recall the *de dicto* beliefs of your previous self.

## 3.5 Rational updating: a more complete procedure

Genuine updating happens in two steps. First you update as if you were in a black box. Then you conditionalize your resulting credences on what you genuinely learn. I have sketched how the first step of updating goes. In order to describe genuine rational updating, I will discuss three ways in which the procedure I sketched is idealized, and how these idealizations can be removed.

### 3.5.1 Credences

So far I have talked about modifying beliefs, rather than credence distributions. But my aim is to develop a general theory of how agents maintain and modify *credences*. Fortunately, an appropriately sophisticated theory of interpersonal communication can again serve as our guide. In making an assertion, you can do much more than simply convey certain *de dicto* beliefs to me. If you say 'John smokes' to me, then I should believe that John smokes. But if you merely say 'John *might* smoke' to me, then you merely propose that I should believe that John might smoke. On some recent theories of modals, this means I should give at least some credence to the proposition that John smokes. Similarly, if you say 'if John smokes, then Mary drinks' to me, then I should give high conditional credence to the proposition that Mary drinks, conditional on the proposition that John smokes. If you say 'it is .9 likely that John smokes' to me, then I should give .9 credence to the proposition that John smokes. By making assertions, you propose that my credences satisfy some constraint, presumably one that your credences already satisfy.[8]

The analogy with updating extends: in black box updating, your current credences should satisfy constraints that your past credences used to satisfy. Earlier I said that *de dicto* beliefs are what you convey in conversation and recall from your previous self. But in fact what you convey and recall are constraints on your credences in *de dicto* propositions. Suppose you used to give .9 credence to a *de se* proposition: that it was the fourth of the month. Given what you used to believe with certainty, this proposition is equivalent with some *de dicto* proposition, to which you also gave .9 credence if your credences were

8. See SWANSON 2006 and YALCIN 2007 for developed theories that relate asserted contents to constraints on credences.

probabilistically coherent. If you are updating in a black box, you should currently give .9 credence to the same *de dicto* proposition.

Black box updating is like communication: it as if your previous self could talk to you and thereby propose constraints on your *de dicto* credences. Only unlike cases of real communication, there is no limit to the amount of information your previous self can convey. It is as if your previous self proposes that your current *de dicto* credences satisfy every constraint that they did before. So in a hypothetical black box updating case, a case where no genuine learning occurs, all of your *de dicto* credences should stay just the same.

### 3.5.2 Conditional credences about your relation to your previous self

So far when talking about how your previous *de dicto* beliefs should influence your current *de se* beliefs, I have talked about your beliefs about your relation to your previous self. But in fact you have more complicated opinions about your relation to your previous self. In particular, your credences about how much time has passed between you and your previous self are conditional in nature. For example, suppose you recently looked at a clock that read 2:00, but you think the clock may be an hour early. Suppose you also know that time passes more quickly as the afternoon wears on. Then you might currently believe that if it was indeed 2:00 earlier, four minutes have passed since you looked at the clock. But if it was 3:00, five or six minutes may have passed. In this way, your opinions about how much time has passed are conditional credences. They are conditional on *de dicto* propositions, such as the *de dicto* proposition you would have used 'now is 2:00' to convey when you were looking at the clock.

In practice, your opinions about your relation to your previous self are given by conditional credence distributions. For any *de dicto* proposition, you have a credence distribution over *de se* propositions, given that *de dicto* proposition. For example, conditional on your having looked at the clock at 3:00, you may give .5 credence to five minutes having passed and .5 credence to six minutes having passed. Conditional credence distributions like these are more precise models of your opinions about your relative location in time.

In black box updating, your credences are entirely determined by two elements: your previous credences in *de dicto* propositions, and your current conditional credences about your relation to your previous self. First your pre-

vious credences determine how much credence you give to any given *de dicto* proposition. Then your conditional credences determine how you distribute that credence among all *de se* propositions entailing that *de dicto* proposition. This uniquely determines a credence distribution over both *de dicto* and *de se* propositions. If your previous opinions and your innate sense of time passing were your only sources of information, your rationally updated credences would be determined in just this way.

### 3.5.3 Genuine learning

Once we understand how you should update in a black box case, describing a complete procedure for rational updating is straightforward. In ordinary cases, your later credences are not only informed by your previous opinions. They must reflect what you genuinely learn as time passes, information that makes you smarter than your previous self. The combination of your previous *de dicto* credences and conditional *de se* credences is a hypothetical credence distribution, representing how you should have updated if you had not genuinely learned anything. In order to arrive at the updated credences you really should have, you must conditionalize this hypothetical credence distribution on what you genuinely learn.[9]

It is important to notice that the first step of updating results in a merely hypothetical credence distribution. For example, it may be that you are always genuinely learning information, so that you never have credences informed only by your own sense of time passing. Rationally updated credences may always be the product of your black box credences and what you genuinely learn.

Distinguishing steps of updating that use different kinds of information allows us to more easily recognize how those steps of updating are related to other processes. The first step of updating is analogous to communication. If you have opinions about how you are related to a speaker, she may convey *de dicto* information that constrains your *de se* credences. If you have opinions about how you are related to your previous self, your previous *de dicto* credences may constrain your current *de se* credences in just the same way. The second

9. This may involve updating by simple conditionalization, Jeffrey conditionalization, or more complicated ways of updating a credence distribution on non-propositional evidence. See Moss 2007 for further discussion.

step of updating is simply conditionalizing on what you learn. In a sense, we have found that conditionalization is the correct procedure for updating *de se* credences. It is just that we must be careful that we are conditionalizing the correct object on what you learn: not your previous credences, but a hypothetical modification of them.

## 3.6 Discussion

I have given a framework that organizes and highlights various features of the updating process. The most dramatic consequence of my framework is that the process of rational updating can be entirely factored into two steps: generating hypothetical credences informed only by your previous opinions and your sense of time passing, and conditionalizing these credences on what you genuinely learn. In other words, two kinds of information inform your later credences. There is information you gain from your innate sense of time passing, and there is genuinely learned information that makes you more informed than your previous self. I have argued that these different kinds of information should play different roles in rational updating.

The framework I have given is more modest than some alternative theories. One respect in which it is modest is that I accept as primitive the distinction between black box updating, and updating as you genuinely learn information. In other words, I accept as primitive the distinction between information you gain from your sense of time passing, and genuinely learned information. I take it that we have some intuitive grasp of this distinction. In order to issue verdicts about particular cases, my theory relies on our intuitive grasp of what counts as information you gain from your sense of time passing.

Other theories generally do not distinguish the kinds of constraints on credences that are inputs to updating. For instance, Titelbaum 2008 calls the inputs to his updating procedure "extrasystematic constraints," and says only that they "represent rational requirements derived from the specific details of the story being modeled" (560). Other theories generally do not recognize information you gain from your sense of time passing as a primitive input to an updating procedure. They simply stipulate that your later self meets certain conditions, such as being certain of the *de se* proposition that it is 5:00, without

saying whether you arrived at this certainty by looking at your watch or by independently keeping track of how much time had passed.

In order to illustrate how my theory works and how it differs from less modest theories, I will conclude by discussing a particular case.[10] Suppose you are being held in prison until Sunday. In prison you lose track of time, so you are unsure whether it is Thursday, Friday, or Saturday. Say you have $1/4$ credence that it is Thursday, $1/4$ credence that it is Friday, and $1/2$ credence that it is Saturday. Suppose that you go to sleep, and immediately upon waking up the next day, you learn that it is not yet Sunday. Intuitively, you should then have $1/2$ credence that it is Friday, and $1/2$ credence that it is Saturday.

The framework I have given yields this verdict. Suppose that instead of waking up to learn that it is not yet Sunday, you wake up in a black box. One day ago, you had $1/4$ credence in the *de dicto* proposition that you would have used 'today is Thursday' to convey, namely that it was Thursday.[11] If you wake up in a black box, you should still have $1/4$ credence in this proposition (§5.1). Furthermore, conditional on the proposition that it was Thursday, you are currently certain that it is Friday. So you must currently have at least $1/4$ credence that it is Friday (§5.2). Similarly, you must have at least $1/4$ credence that it is Saturday, and $1/2$ credence that it is Sunday. So if you wake up in a black box, your credences about what day it is should simply be shifted forward by one day. In the real prison case, you should update by conditionalizing these shifted credences on what you genuinely learn when you wake up: that it is not Sunday (§5.3). Hence my framework confirms our intuition that on waking up in the prison case, you should have $1/2$ credence that it is Friday and $1/2$ credence that it is Saturday.

Other theories have more trouble yielding this verdict. TITELBAUM 2008 gives a theory that is similar to mine in some respects. But the prison case presents a problem for Titelbaum. This problem is useful: it distinguishes Titelbaum's theory from mine, and highlights the importance of distinguishing the kinds of constraints on credences that are inputs to updating.

10. See ARNTZENIUS 2003 and BRADLEY 2008 for structurally similar examples.

11. I adopt the following conventions throughout this section: 'that it is Thursday' refers to a *de se* proposition, namely the set of centered worlds centered on Thursday, and 'that it was Thursday' refers to the *de dicto* proposition you would have used 'today is Thursday' to convey before waking up.

Titelbaum argues that rational updating is governed by a pair of principles.[12] The first principle relates credence distributions that an agent can have over different algebras. Suppose there is some algebra $P$ of propositions over which your credences are defined at various times, and you add propositions to $P$ to get a larger algebra $P'$. Suppose your extension of $P$ is conservative, in the following sense: you add only propositions $p'$ such that at every time, you can find some proposition $p$ already in $P$ such that you are certain that $p'$ and $p$ have the same truth value. In this case, we say that $P'$ is a *proper expansion* of $P$ for you.

Titelbaum's first principle relates how you update your credence distributions over an algebra and its proper expansion:

(PE) If rationality constrains your credence distribution over $P$, and $P'$ is a proper expansion of $P$ for you, then update your credence distribution over $P'$ according to the same constraints.

In other words, your updated $P'$ credences must satisfy any constraint your updated $P$ credences must satisfy.

Titelbaum's second principle is a limited version of conditionalization. Titelbaum states one version of the principle, and later qualifies the principle in response to counterexamples. Here is the unqualified version:

(LC) If you do not become uncertain of any proposition in $P$, then update your credence distribution over $P$ by conditionalizing on the strongest proposition in $P$ of which you become certain.

Titelbaum characterizes (LC) as "conditionalization with one added condition: (LC) relates credences at two times only when the earlier time's certainty set is a subset of the later's" (568). (LC) is designed to avoid standard problems for unrestricted conditionalization. For instance, (LC) does not entail that if you are ever certain that it is 3:00pm, you must always remain certain that it is 3:00pm. Titelbaum says that your experience may directly constrain your credences so that you become uncertain that it is 3:00pm. In that case, the antecedent of

12. I present a simplified version of the theory in TITELBAUM 2008. For instance, Titelbaum says credences are defined on sentences, rather than sets of centered worlds. The simplifying assumptions I make do not affect my arguments against Titelbaum.

(LC) fails, so (LC) does not constrain how you update your credences in *de se* propositions.

Together, (PE) and (LC) entail extensive constraints on how you update your credences. (LC) entails rational constraints on how you update your credence distribution over certain algebras. (PE) then entails further constraints on how you update your credence distribution over proper extensions of those algebras. The idea behind (LC) is that we should refrain from applying conditionalization in just those cases where its application gets us into trouble: cases where we lose certainty in propositions. In such cases, (PE) will tell us how to update our credences.

But as Titelbaum recognizes, even his limited principle of conditionalization entails too much. In particular, (LC) entails counterintuitive verdicts about the prison case. In the prison case, your credences are defined over an algebra generated by the following *de se* propositions:

(P1) that it is Thursday, that it is Friday, that it is Saturday, that it is Sunday

Remember that when you wake up in the prison case, you immediately learn that it is not Sunday. So you do not become uncertain of any (P1) proposition. Before you go to sleep, you are certain that it is Thursday, Friday, or Saturday. On waking and learning that it is not Sunday, you are again certain of the same *de se* proposition. Since you do not become uncertain of any (P1) proposition, the unqualified version of (LC) says that when you wake up, you should update your (P1) credences by conditionalizing on the strongest (P1) proposition you learn: that it is not Thursday. But if you updated this way, you would end up with the wrong credences: $1/3$ credence that it is Friday, and $2/3$ credence that it is Saturday.

The idea behind this counterexample to (LC) is that applying conditionalization can get us into trouble, even when we do not lose certainty in propositions. Preserving certainties is just one bad effect of updating *de se* credences by conditionalization. Failing to appropriately shift ratios of credences in *de se* propositions is another.

In order to avoid a bad verdict about the prison case, Titelbaum introduces

one final principle: a "modeling rule" (579). The modeling rule qualifies (LC) as follows:

> If we have a model and its expansion, and the analogues of the model's verdicts are not verdicts of the expansion, we should not trust the original model's verdicts to represent requirements of ideal rationality... we should trust the verdicts of the model whose language is a superset of the languages of all the models we have tried. (579)

In other words, Titelbaum really only endorses weaker versions of (LC) and (PE). The weakened principles are such that if they entail some claim about your credences over an algebra, they entail the same claim about your credences over every extension of that algebra.[13]

In particular, (LC) as initially stated always tells you how to update your credence distribution over *de dicto* propositions when you do not become uncertain of any *de dicto* propositions. But the weakened version of (LC) does not. Once we weaken (LC), it entails constraints on how you update your credence distribution over a given algebra only when other principles entail the same constraints on how you update your credence distribution over any extension of the given algebra. For instance, (LC) constrains how you update your credence distribution over *de dicto* propositions only if other principles similarly constrain how you update your credence distribution over both *de dicto* and *de se* propositions.

Titelbaum anticipates that qualifying (LC) in this way will keep (LC) from entailing counterintuitive consequences. He discusses his fix for (LC) in the following passage:

> Suppose that in some model $M$ (LC) applies to yield diachronic verdicts. But suppose an improper expansion of $M$, $M^+$, represents context-sensitive claims in its modeling language that are not represented in the modeling language of $M$. If any of these extra claims goes from certainty to less-than-certainty during the story, (LC) will fail to yield any diachronic verdicts for $M^+$... [and] because [$M^+$] fails to replicate $M$'s diachronic verdicts we

13. Presumably Titelbaum wants to endorse the strongest weakened versions of (LC) and (PE) satisfying this condition. I do not see why there should be a unique way to weaken the principles, or a unique set of strongest weakened principles satisfying the modeling rule. Titelbaum does not address these questions.

> should not rely on those verdicts to represent requirements of ideal rationality. (18)

In other words, Titelbaum says that the qualified version of (LC) will not yield counterintuitive verdicts about updating your credences on small algebras, since such verdicts will fail to be entailed by principles about how you should update your credences over extensions of such algebras.

I agree with Titelbaum that in some cases, the modeling rule prevents (LC) from entailing some counterintuitive consequences. For instance, qualifying (LC) with the modeling rule forestalls a counterintuitive verdict in the prison case.[14] In the prison case, you start to have non-zero credence in a certain *de se* proposition: roughly, that one day has passed between you and your previous self. So you become uncertain of the negation of this proposition: that one day has not passed. In addition to (P1), your credences are defined over the algebra generated by the following propositions:

(P2) that it is Thursday, that it is Friday, that it is Saturday, that it is Sunday, that one day has not passed

Since you become uncertain of a proposition in (P2), (LC) as applied to (P2) does not entail that your later credence that it is Friday should be $1/3$. There simply are no principles that entail that you should have $1/3$ credence that it is Friday when you update your credence distribution over (P2). So by the modeling rule, this claim is not entailed by the qualified version of (LC) as applied to your credence distribution over the smaller algebra (P1). The qualified version of (LC) does not entail that your later credence that it is Friday should be $1/3$.

So far, so good. But unfortunately, the prison case raises a further problem for (LC). Once qualified, (LC) does not entail a counterintuitive verdict about the prison case. But (LC) does not entail the correct verdict about the prison case either. Titelbaum still needs to derive the conclusion that on waking, you should be equally confident that it is Friday and that it is Saturday.

14. See p.19 of TITELBAUM 2008 for his discussion of cases like the prison case. In personal communication, Titelbaum confirms that this is his intended response.

Titelbaum does not address this problem in detail. But I take it that the following remarks contain his proposed solution:

> If the modeling language of $M^+$... contains context-insensitive truth-value equivalents for each context-sensitive sentence at each time in the time set – we can construct a reduction of $M^+$ different from $M$ whose language represents only the context-insensitive claims represented in $M^+$. This model will yield diachronic verdicts by (LC), and since it is a proper reduction of $M^+$ those verdicts can be exported back to $M^+$ by (PE). (579)

Given these remarks, here is my reconstruction of how Titelbaum aims to derive the correct verdict about how you should update in the prison case. In the prison case, you have credences in the (P2) propositions. But you also have credences in *de dicto* propositions, such as the proposition that your previous self would have used 'today is Thursday' to convey, i.e. that it was Thursday. So your credences are also defined over algebras generated by the following propositions:

(P3) that it is Thursday, that it is Friday, that it is Saturday, that it is Sunday, that one day has not passed, that it was Thursday, that it was Friday, that it was Saturday

(P4) that it was Thursday, that it was Friday, that it was Saturday

Before you go to sleep, you are certain that the proposition that one day has not passed is true. Once you wake up, you are certain that it is false. So at each time, you are certain that the proposition that one day has passed is equivalent to some particular proposition in both (P1) and (P4), namely either the set of all worlds or the set of no worlds. Furthermore, you can always relate the *de se* propositions in (P1) and (P3) to the *de dicto* propositions in (P4) and (P3). For instance, when you first wake up you are certain that the proposition that it was Thursday when you went to sleep is true just in case it is Friday. So we can conclude that (P3) is a proper expansion of both (P1) and (P4).

It appears that Titelbaum could use these facts to derive the correct verdict about the prison case, namely that you should update your credence distribution over (P1) by coming to have $\frac{1}{2}$ credence that it is Friday.[15] The derivation

15. Following the passage quoted above, let $M$ be (P1), let $M^+$ be (P3), and let the "reduction of

proceeds in three steps. First, note that on waking, you do not become uncertain of any proposition in (P4). So (LC) says that you should update your credence distribution over (P4) by conditionalizing on the information you learn, namely that either it was Thursday or it was Friday when you went to sleep. On the first day, you started with $1/4$ credence that it was Thursday and $1/4$ credence that it was Friday. So after conditionalizing, you should have $1/2$ credence that it was Thursday. Second, since (P3) is a proper expansion of (P4), (PE) says that you should update your credence distribution over (P3) in the same way, by coming to have $1/2$ credence that it was Thursday, and $1/2$ credence that it is Friday. Finally, recall the modeling rule: "ultimately we should trust the verdicts of the model whose language is a superset of the languages of all the models we have tried for a story" (580). This modeling rule entails that since (P3) is an expansion of (P1), you should also update your credence distribution over (P1) by coming to have $1/2$ credence that it was Thursday, and $1/2$ credence that it is Friday.

The problem with this chain of reasoning is that it works in both directions. If the above reasoning is acceptable, then we should also be able to reason as follows: by (LC), you should update your credence distribution over (P1) by coming to have $1/3$ credence that it is Friday. By (PE), you should update your credence distribution over (P3) in the same way. So you should update your credence distribution over (P3) by coming to have $1/3$ credence that it was Thursday. By the modeling rule, you should update your credence distribution over (P4) in the same way. Hence you should update your credence distribution over (P4) by coming to have $1/3$ credence that it was Thursday and $2/3$ credence that it was Friday, rather than becoming equally confident of these propositions.

As far as Titelbaum's theory is concerned, the algebras (P1) and (P4) stand in symmetric relations to the larger algebra (P3). So Titelbaum faces a dilemma. If (LC) applies to your credence distribution over (P4), then it should apply to your credence distribution over (P1), and (LC) will yield the wrong verdict about how to update that credence distribution. But if (LC) does not apply to your credence distribution over (P4), then Titelbaum cannot use its verdict about (P4) to derive the right verdict about how to update your credence distribution over (P1).

---

$M^+$ different from $M$ whose language represents only the context-insensitive claims represented in $M^+$" be (P4).

Titelbaum has not demonstrated that he can prevent (LC) from generating counterintuitive consequences, without also preventing (LC) from generating the right verdicts about updating in particular cases. This problem with his theory is very wide in scope. In any natural case of updating, you will lose your certainty that some small amount of time has not passed. So the unqualified version of (LC) will not apply to your entire credence distribution. Instead we must always figure out how constraints on coarser credences to which (LC) applies will induce constraints on your real credences. If we do not know which coarser credences are relevant to constraining your real credences, there will be no way to say what your real credences should be.

This problem for Titelbaum's theory highlights an important feature of rational updating. Titelbaum has trouble with the prison case because nothing in his theory distinguishes the algebra (P1) containing *de se* propositions about what day it is, and the algebra (P4) containing *de dicto* propositions about what day it was on a particular occasion. The prison case illustrates that your opinions about these propositions play different roles in updating. Intuitively, your current credences about what day it is should be informed by your previous credences about what day it was, and by your current information that it is not Sunday. This means that your opinion that one day has passed is not on a par with your information that it is not Sunday. Instead your opinion about how much time has passed plays a special role, namely determining how your previous (P4) credences inform your current (P1) credences.

The moral of the prison case is that a theory of updating should distinguish your credences about what day it was, what day it is, and how much time has passed. This is what my framework does. Once we recognize that different credences inform your updated credence distribution in different ways, our theory naturally yields intuitively correct verdicts about cases of rational updating.

## References


ALSTON, W. 2005. *Beyond "Justification": Dimensions of Epistemic Evaluation.* Cornell University Press, Ithaca.

ANAND, PRANAV. 2006. "De *De Se.*" PhD. dissertation, Department of Linguistics and Philosophy, MIT.

ARNTZENIUS, FRANK. 2003. "Some Problems for Conditionalization and Reflection." *Journal of Philosophy*, vol. 100: 356–70.

BRADLEY, DARREN. 2008. "How Belief Mutation Saves Conditionalization from Self-Locating Information." Ms., Department of Philosophy, University of British Columbia. Available at `http://faculty.arts.ubc.ca/dbradley/`.

BRIER, GLENN W. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review*, vol. 78 (1): 1–3.

CHIERCHIA, GENNARO. 1989. "Anaphora and Attitudes De Se." In *Semantics and Contextual Expression*, R. BARTSCH, J. VAN BENTHEM & VAN EMDE BOAS, editors, 1–31. Foris, Dordrecht.

CHRISTENSEN, DAVID. 2007. "Epistemology of Disagreement: the Good News." *Philosophical Review.*

DEROSE, KEITH. 1994. "Lewis on 'Might' and 'Would' Counterfactual Conditionals." *Canadian Journal of Philosophy*, vol. 24 (3): 413–418.

—. 1997. "Can It Be That It Would Have Been Even Though It Might Not Have Been?" In *Philosophical Perspectives 11: Mind, Causation, and World*, JAMES TOMBERLIN, editor, 385–413. Blackwell Publishers, Ltd., Oxford.

DUMMETT, MICHAEL. 1991. *The Logical Basis of Metaphysics.* Duckworth, London.

EDGINGTON, DOROTHY. 2003. "Counterfactuals and the Benefit of Hindsight." In *Causation and Counterfactuals*, P. DOWE & P. NOORDHOF, editors, 12–27. Routledge, London.

—. 2008. "Do Counterfactuals Have Truth Conditions?" Ms., Birkbeck College, London.

EELLS, ELLERY. 1982. *Rational Decision and Causality.* Cambridge University Press, Cambridge.

EGAN, ANDY. 2005. "Epistemic Modals, Relativism, and Assertion." *Philosophical Studies*, vol. 133 (1): 1–22.

ELGA, ADAM. 2006. "Reflection and Disagreement." Forthcoming in *Nous*.

—. 2007. "Subjective Probabilities Should Be Sharp." Ms., Princeton University.

VON FINTEL, KAI. 1997. "Bare Plurals, Bare Conditionals, and *Only*." *Journal of Semantics*, vol. 14: 1–56.

—. 2001. "Counterfactuals in a Dynamic Context." In *Ken Hale: A Life in Language*, MICHAEL KENSTOWICZ, editor. MIT Press, Cambridge.

—. 2005. "LSA 311: Lecture 9." Handout from LSA 2005, *Pragmatics in Linguistic Theory*, URL `http://semantics-online.org/lsa311/lsa311-ho-9.pdf`.

FITELSON, BRANDON & DAVID JEHLE. 2007. "What is the 'Equal Weight View'?" Ms., University of California, Berkeley. Available at `http://fitelson.org/ew.pdf`.

FODOR, JANET DEAN. 1970. "The Linguistic Description of Opaque Contexts." PhD. dissertation, Department of Linguistics and Philosophy, MIT.

VAN FRAASSEN, BAS C. 1980. "A Temporal Framework for Conditionals and Chance." *Philosophical Review*, vol. 89: 91–108.

—. 2006. "Vague Expectation Value Loss." *Philosophical Studies*, vol. 127: 483–491.

FRENCH, S. 1985. "Group Consensus Probability Distributions: A Critical Survey." In *Bayesian statistics 2*, J. M. BERNARDO, M. H. DEGROOT, D. V. LINDLEY & A. F. M. SMITH, editors, 183–197. North-Holland, Amsterdam.

GENEST, CHRISTIAN & JAMES ZIDEK. 1986. "Combining Probability Distributions: A Critique and an Annotated Bibliography." *Statistical Science*, vol. 1: 114–135.

GIBBARD, ALLAN. 2006. "Rational Credence and the Value of Truth." In *Oxford Studies in Epistemology*, TAMAR SZABÓ GENDLER & JOHN HAWTHORNE, editors, vol. 2. Oxford University Press, Oxford.

GIBBARD, ALLAN & WILLIAM L. HARPER. 1978. "Counterfactuals and Two Kinds of Expected Utility." In *Foundations and Applications of Decision Theory*, C. HOOKER, J. LEACH & E. MCCLENNEN, editors, 123–62. Reidel, Dordrecht.

GILLIES, THONY. 2007. "Counterfactual Scorekeeping." *Linguistics and Philosophy*, vol. 30: 329–360.

GOOD, I. J. 1952. "Rational Decisions." *Journal of the Royal Statistical Society, Series B*, vol. 14 (1): 107–114.

HÁJEK, ALAN. 2007. "Most Counterfactuals are False." Ms., Australian National University:
`http://philrsss.anu.edu.au/people-defaults/alanh/papers/MCF.pdf`.

HARPER, WILLIAM L., ROBERT STALNAKER & GLENN PEARCE, editors. 1981. *Ifs: Conditionals, Belief, Decision, Chance, and Time*. D. Reidel Publishing Company, Dordrecht.

HELLER, MARK. 1995. "Might-Counterfactuals and Gratuitous Differences." *Australasian Journal of Philosophy*, vol. 73: 91–101.

JEFFREY, RICHARD C. 1968. "Probable Knowledge." In *Probability and the Art of Judgment*, 30–43. Cambridge University Press, Cambridge.

—. 1983. "Bayesianism With a Human Face." In *Testing Scientific Theories*, JOHN EARMAN, editor. University of Minnesota.

—. 1986. "Probabilism and Induction." *Topoi*, vol. 5: 51–58.

JOYCE, JAMES M. 1998. "A Nonpragmatic Vindication of Probabilism." *Philosophy of Science*, vol. 65 (4): 575–603.

—. 2005. "How Probabilities Reflect Evidence." *Philosophical Perspectives*, vol. 19: 153–178.

—. 2007. "Epistemic Deference: The Case of Chance." *Proceedings of the Aristotelian Society*, vol. 107 (1): 187–206.

KAPLAN, DAVID. 1989. "Afterthoughts." In *Themes from Kaplan*, JOSEPH ALMOG, JOHN PERRY & HOWARD WETTSTEIN, editors, 565–614. Oxford University Press, Oxford.

KAPLAN, MARK. 1996. *Decision Theory as Philosophy*. Cambridge University Press, Cambridge.

KARTUNNEN, LAURIE & STANLEY PETERS. 1976. "What Indirect Questions Conventionally Implicate." In *Papers from the 12th Regional Meeting of the Chicago Linguistics Society*, C. WALKER S. MUFWENE & S. STEEVER, editors. Chicago Linguistics Society, Chicago.

KATSUNO, HIROFUMI & ALBERTO MENDELZON. 1991. "On the Difference Between Updating a Knowledge Base and Revising it." *Proceedings of the 2nd Principles of Knowledge Representation and Reasoning Conference*, 387–394.

KELLY, TOM. 2007. "Peer Disagreement and Higher Order Evidence." Manuscript, Princeton University. Available at `http://www.princeton.edu/~tkelly/papers.htm`.

KRIFKA, MANFRED. 1996. "Pragmatic Strengthening in Plural Predications and Donkey Sentences." CLC Publications, Ithaca.

LAKOFF, GEORGE. 1972. "Linguistics and Natural Logic." In *Semantics of Natural Language*, DONALD DAVIDSON & GILBERT HARMAN, editors, 545–665. Reidel, Dordrecht.

LEVI, ISAAC. 1980. *The Enterprise of Knowledge*. MIT Press, Cambridge.

LEWIS, DAVID K. 1973a. *Counterfactuals*. Basil Blackwell Ltd., Malden, MA.

—. 1973b. "Counterfactuals and Comparative Possibility." In HARPER et al. (1981).

—. 1979. "Attitudes *De Dicto* and *De Se*." *Philosophical Review*, vol. 88: 513–43.

—. 1980. "A Subjectivist's Guide to Objective Chance." In *Philosophical Papers*, vol. 2, 83–132. Oxford University Press, Oxford. With postscript.

—. 1999. *Papers in Metaphysics and Epistemology*. Cambridge University Press, Cambridge.

LÖBNER, SEBASTIAN. 1987. "The Conceptual Nature of Natural Language Quantification." In *Proceedings of the '87 Debrecen Symposium on Logic and Language*, I. RUSZA & A. SZABOLCSI, editors. Akademiai Kaido, Budapest.

LOWE, E. J. 1995. "The Truth about Counterfactuals." *Philosophical Quarterly*, vol. 45: 41–59.

LYNCH, M. forthcoming. "The Values of Truth and the Truth of Values." In *Epistemic Value*, A. HADDOCK, A. MILLAR & D. H. PRITCHARD, editors. Oxford University Press, Oxford.

MORGAN, JERRY. 1970. "On the Criterion of Identity for Noun Phrase Deletion." *Chicago Linguistics Society*, vol. 6.

MOSS, SARAH. 2007. "Non-Propositional Factives." Ms., Department of Linguistics and Philosophy, MIT.

NINAN, DILIP. 2008. "Imagination, Content, and the Self." PhD. dissertation, Department of Linguistics and Philosophy, MIT.

—. 2009. "*De Se* Attitudes: Ascription and Communication." Ms., Department of Philosophy, St. Andrews.

ODDIE, GRAHAM. 1997. "Conditionalization, Cogency, and Cognitive Value." *British Journal for the Philosophy of Science*, vol. 48 (4): 533–41.

PERCIVAL, PHILIP. 2002. "Epistemic Consequentialism." *Proceedings of the Aristotelian Society*, vol. 76 (1): 121–51.

POSTAL, PAUL M. 1971. *Cross-Over Phenomena*. Holt, Rinehart, and Winston, New York.

QUINE, W. V. O. 1950. *Methods of Logic*. Henry Holt and Company, New York.

—. 1969. "Propositional Objects." In *Ontological Relativity and Other Essays*, 139–160. Columbia University Press, New York.

RAMSEY, F. P. 1931. "General Propositions and Causality." Routledge & Kegan Paul, London.

SAVAGE, LEONARD. 1971. "Elicitation of Personal Probabilities and Expectations." *Journal of the American Statistical Association*, vol. 66: 783–801.

SCHEIN, BARRY. 2001. "Adverbial, Descriptive Reciprocals." CLC Publications, Ithaca.

SCHLENKER, PHILIPPE. 2004. "Conditionals as Definite Descriptions." *Research on Language and Computation*, vol. 2 (3): 417–62.

SHOGENJI, TOMOJI. 2007. "A Conundrum in Bayesian Epistemology of Disagreement." Ms., Rhode Island College. Available at `http://socrates.berkeley.edu/~fitelson/few/few_07/shogenji.pdf`.

SKYRMS, BRYAN. 1978. "The Prior Propensity Account of Subjunctive Conditionals." In HARPER et al. (1981).

SLOTE, M. 1978. "Time in Counterfactuals." *Philosophical Review*, vol. 87: 3–27.

SMITH, CEDRIC. 1961. "Consistency in Statistical Inference and Decision." *Journal of the Royal Statistical Society Series B*, vol. 23: 1–37.

STALNAKER, ROBERT C. 1968. "A Theory of Conditionals." In HARPER et al. (1981), 41–55.

—. 1978. "A Defense of Conditional Excluded Middle." In HARPER et al. (1981), 41–55.

—. 2008. "Locating Ourselves in the World." Ms., Dept. of Linguistics and Philosophy, MIT.

STANLEY, JASON. 1997. "Names and Rigid Designation." 555–585. Blackwell Publishers, Ltd., Oxford.

VON STECHOW, ARNIM. 2002. "Binding by Verbs: Tense, Person and Mood under Attitudes."

SWANSON, ERIC. 2006. "Interactions with Context." PhD. dissertation, Department of Linguistics and Philosophy, MIT.

TELLER, PAUL. 1973. "Conditionalization and Observation." *Synthese*, vol. 26: 218–258.

TINTNER, GERHARD. 1941. "The Theory of Choice Under Subjective Risk and Uncertainty." *Econometrica*, vol. 9: 298–304.

TITELBAUM, MICHAEL. 2008. "The Relevance of Self-Locating Beliefs." *Philosophical Review*, vol. 117 (4): 555–606.

WALLEY, PETER. 1991. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London.

WARMBROD, KEN. 1981. "Counterfactuals and Substitution of Equivalent Antecedents." *Journal of Philosophical Logic*, vol. 10: 267–289.

WEATHERSON, BRIAN. 2007. "Disagreeing about Disagreement." Ms., Department of Philosophy, Cornell University. Available at `http://brian.weatherson.org/DaD.pdf`.

WILLIAMS, P. M. 1976. "Indeterminate Probabilities." In *Formal Methods in the Methodology of Empirical Sciences*, M. PRZELECKI, K. SZANIAWSKI & R. WOJCICKI, editors, 229–246. Reidel, Dordrecht.

WILLIAMS, ROBBIE. 2006a. "Conversation and Conditionals." Forthcoming in *Philosophical Studies*.

—. 2006b. "There is no 'might' argument against Conditional Excluded Middle." Ms., University of Leeds: http://www.personal.leeds.ac.uk/~phljrgw/wip/mightCEM.pdf.

WILLIAMSON, TIMOTHY. 1994. *Vagueness.* Routledge, London.

YALCIN, SETH. 2007. "Epistemic Modals." *Mind*, vol. 116: 983–1026.